

Volume 13 Issue 5

May 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 5 May 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

Doroła 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: Implementation of Data Mining on a Secure Cloud Computing over a Web API using Supervised Machine Learning Algorithm

Authors: Tosin Ige, Sikiru Adewale

PAGE 1 – 4

Paper 2: AI Powered Anti-Cyber Bullying System using Machine Learning Algorithm of Multinomial Naïve Bayes and Optimized Linear Support Vector Machine

Authors: Tosin Ige, Sikiru Adewale

PAGE 5 – 9

Paper 3: Replica Scheduling Strategy for Streaming Data Mining

Authors: Shufan Li, Siyuan Yu, Fang Xiao

PAGE 10 – 19

Paper 4: Use of Neural Networks in the Adaptive Testing System

Authors: Ekaterina Vitalevna Chumakova, Tatiana Alexandrovna Chernova, Yulia Aleksandrovna Belyaeva, Dmitry Gennadievich Korneev, Mikhail Samuilovich Gasparian

PAGE 20 – 27

Paper 5: Development of Ontology-based Domain Knowledge Model for IT Domain in e-Tutor Systems

Authors: Ghanim Hussein Ali Ahmed, Jawad Alshboul, Laszlo Kovacs

PAGE 28 – 34

Paper 6: The Design of Home Fire Monitoring System based on NB-IoT

Authors: Jun Wang, Ting Ke, Mengjie Hou, Gangyu Hu

PAGE 35 – 42

Paper 7: The Effect of Natural Language Processing on the Analysis of Unstructured Text: A Systematic Review

Authors: Walter Luis Roldan-Baluis, Noel Alcas Zapata, Maria Soledad Manaccasa Vasquez

PAGE 43 – 51

Paper 8: Hybrid Fault Diagnosis Method based on Wavelet Packet Energy Spectrum and SSA-SVM

Authors: Jinglei Qu, Bingxin Ma, Xiaojie Ma, Mengmeng Wang

PAGE 52 – 60

Paper 9: Multi-instance Finger Vein-based Authentication with Secured Templates

Authors: Swati K. Choudhary, Ameya K. Naik

PAGE 61 – 71

Paper 10: SHD-IoV: Secure Handover Decision in IoV

Authors: Hala E. I. Jubara

PAGE 72 – 80

Paper 11: The Impact of Security and Payment Method On Consumers' Perception of Marketplace in Saudi Arabia

Authors: Mdawi Alqahtani, Marwan Ali Albahar

PAGE 81 – 88

Paper 12: Individual Risk Classification of Crime Groups using Ensemble Classifier Method

Authors: Ardhito P. Anggana, Amalia Zahra

PAGE 89 – 98

Paper 13: Evaluating the Effectiveness and Usability of AR-based OSH Application: HazHunt

Authors: Ahmad A. Kamal, Syahrul N. Junaini, Abdul H. Hashim

PAGE 99 – 106

Paper 14: Efficacy of the Image Augmentation Method using CNN Transfer Learning in Identification of Timber Defect

Authors: Teo Hong Chun, Ummi Rabaah Hashim, Sabrina Ahmad, Lizawati Salahuddin, Ngo Hea Choon, Kasturi Kanchymalay

PAGE 107 – 114

Paper 15: Security Analysis on an Improved Anonymous Authentication Protocol for Wearable Health Monitoring Systems

Authors: Gayeong Eom, Haewon Byeon, Younsung Choi

PAGE 115 – 121

Paper 16: A Review on Bio-inspired Optimization Method for Supervised Feature Selection

Authors: Montha Petwan, Ku Ruhana Ku-Mahamud

PAGE 122 – 132

Paper 17: Implementation of a Data Protection Model dubbed Harricent_RSECC

Authors: Frimpong Twum, Vincent Amankona, Yaw Marfo Missah, Ussiph Najim, Michael Opoku

PAGE 133 – 145

Paper 18: Goal Question Metric as an Interdisciplinary Tool for Assessing Mobile Learning Application

Authors: Sim Yee Wai, Cheah WaiShiang, Piau Phang, Kai-Chee Lam, Eaqerzilla Phang, Nurfauzia binti Jali

PAGE 146 – 152

Paper 19: Spatial Feature Fusion for Biomedical Image Classification based on Ensemble Deep CNN and Transfer Learning

Authors: Sanskruti Patel, Rachana Patel, Nilay Ganatra, Atul Patel

PAGE 153 – 159

Paper 20: Effectivity Score of Simulation Tools towards Modelling Design in Internet-of-Things

Authors: Gauri Sameer Rapate, N C Naveen

PAGE 160 – 170

Paper 21: Analysis and Prediction of COVID-19 by using Recurrent LSTM Neural Network Model in Machine Learning

Authors: N. P. Dharani, Polaiah Bojja

PAGE 171 – 178

Paper 22: A Proposed Fraud Detection Model based on e-Payments Attributes a Case Study in Egyptian e-Payment Gateway

Authors: Mohamed Hassan Nasr, Mohamed Hassan Farrag, Mona Mohamed Nasr

PAGE 179 – 186

Paper 23: Improved ISODATA Clustering Method with Parameter Estimation based on Genetic Algorithm

Authors: Kohei Arai

PAGE 187 – 193

Paper 24: A Pre-trained Neural Network to Predict Alzheimer's Disease at an Early Stage

Authors: Ragavamsi Davuluri, Ragupathy Rengaswamy

PAGE 194 – 200

Paper 25: A k-interpolation Model Clustering Algorithm based on Kriging Method

Authors: Guoyan Chen, Yaping Qian

PAGE 201 – 207

Paper 26: A Food Waste Mobile Gamified Application Design Model using UX Agile Approach in Malaysia

Authors: Nooralisa Mohd Tuah, Siti Khadijah Abd. Ghani, Suryani Darham, Suaini Sura

PAGE 208 – 217

Paper 27: Entanglement Quantification and Classification: A Systematic Literature Review

Authors: Amirul Asyraf Zahir, Siti Munirah Mohd, Mohd Ilias M Shuhud, Bahari Idrus, Hishamuddin Zainuddin, Nurhidaya Mohamad Jan, Mohamed Ridza Wahiddin

PAGE 218 – 225

Paper 28: Effective Cross Synthesized Methodology for Movie Recommendation with Emotion Analysis through Ranking Score

Authors: R Lavanya, B Bharathi

PAGE 226 – 233

Paper 29: Supervisory Control and Data Acquisition System for Machines used for Thermal Processing of Materials

Authors: Diego Patino, Wilson Tafur Preciado, Albert Miyer Suarez Castrillon, Sir-Alexci Suarez Castrillon

PAGE 234 – 240

Paper 30: Modeling Wireless Mesh Networks for Load Management

Authors: Soma Pandey, Govind R. Kadambi

PAGE 241 – 251

Paper 31: A Model for Classification and Diagnosis of Skin Disease using Machine Learning and Image Processing Techniques

Authors: Shaden Abdulaziz AlDera, Mohamed Tahar Ben Othman

PAGE 252 – 259

Paper 32: A Hybrid Material Generation Algorithm with Probabilistic Neural Networks for Solving Classification Problems

Authors: Mohammad Wedyan, Omar Alshaweesh, Enas Ramadan, Ryan Alturki, Foziah Gazzawe, Mohammed J. Alghamdi

PAGE 260 – 266

Paper 33: IoT based Portable Weather Station for Irrigation Management using Real-Time Parameters

Authors: Geeta Ambildhuke, Barnali Gupta Banik

PAGE 267 – 278

Paper 34: E-Evaluation based on CSE-UCLA Model Refers to Glickman Pattern for Evaluating the Leadership Training Program

Authors: Ketut Rasmulyani, I Made Yudana, I Nyoman Natajaya, Dewa Gede Hendra Divayana

PAGE 279 – 294

Paper 35: Rule-based Text Extraction for Multimodal Knowledge Graph

Authors: Idza Aisara Norabid, Fariza Fauzi

PAGE 295 – 304

Paper 36: Soil Color as a Measurement for Estimation of Fertility using Deep Learning Techniques

Authors: N Lakshmi Kalyani, Kolla Bhanu Prakash

PAGE 305 – 310

Paper 37: A Survey on Genomic Dataset for Predicting the DNA Abnormalities Using MI

Authors: Siripuri Divya, Y. Bhavani, Thota Mahesh Kumar

PAGE 311 – 319

Paper 38: An Architecture of Domain Independent and Extensible Intelligent Tutoring System based on Concept Dependencies and Subject Paths

Authors: Sanjay Singh, Vikram Singh

PAGE 320 – 329

Paper 39: Secure Routing Protocol for Low Power and Lossy Networks Against Rank Attack: A Systematic Review

Authors: Laila Al-Qaisi, Suhaidi Hassan, Nur Haryani Binti Zakaria

PAGE 330 – 339

Paper 40: Validation of Evacuation Assessment Algorithm in Finding the Best Indoor Evacuation Model

Authors: Amir Haikal Abdul Halim, Khyrina Airin Fariza Abu Samah

PAGE 340 – 347

Paper 41: Effective Prediction of Software Defects using Random-tree Entropy based Feature Selection Framework

Authors: Abdulaziz Alhumam

PAGE 348 – 354

Paper 42: Parameter Optimization of Nonlinear Piezoelectric Energy Harvesting System for IoT Applications

Authors: Li Wah Thong, Swee Leong Kok, Roszaidi Ramlan

PAGE 355 – 363

Paper 43: Smart Blended Learning Framework based on Artificial Intelligence using MobileNet Single Shot Detector and Centroid Tracking Algorithm

Authors: Abdul Wahid, Muhammad Fajar B, Jumadi M. Parenreng, Seny Luhriyani, Puput Dani Prasetyo Adi

PAGE 364 – 369

Paper 44: Smart Agriculture Monitoring System using Clean Energy

Authors: Karim ABOUELMEHDI, Kamal ELHATTAB, Abdelmajid EL MOUTAOUAKKIL

PAGE 370 – 377

Paper 45: RENTAKA: A Novel Machine Learning Framework for Crypto-Ransomware Pre-encryption Detection

Authors: Wira Z. A. Zakaria, Mohd Faizal Abdollah, Othman Mohd, S. M. Warusia Mohamed S. M. M Yassin, Aswami Ariffin

PAGE 378 – 385

Paper 46: Non-contact Facial based Vital Sign Estimation using Convolutional Neural Network Approach

Authors: Nor Surayahani Suriani, Nur Syahida Shahdan, Nan Md. Sahar, Nik Shahidah Affi Md. Taujuddin

PAGE 386 – 393

Paper 47: A Survey on MCT vs. DCT: Who is the Winner in COVID-19

Authors: Omar Khaftab

PAGE 394 – 400

Paper 48: Social Customer Relationship Management as a Communication Tool for Academic Communities in Higher Education Institutions through Social Media

Authors: Ali Ibrahim, Ermatita, Saparudin

PAGE 401 – 411

Paper 49: Application of the Clahe Method Contrast Enhancement of X-Ray Images

Authors: Omarova G. S, Starovoitov V. V, Aitkozha Zh. Zh, Bekbolatov S, Ostayeva A. B, Nuridinov O

PAGE 412 – 420

Paper 50: Efficient Segment-based Image Ciphering using Discretized Chaotic Standard Map with ECB, OFB and CBC

Authors: Mohammed A. AlZain

PAGE 421 – 428

Paper 51: Flower Pollination Algorithm for Feature Selection in Tweets Sentiment Analysis

Authors: Muhammad Iqbal Abu Latiffi, Mohd Ridzwan Yaakub, Ibrahim Said Ahmad

PAGE 429 – 436

Paper 52: Re-CRUD Code Automation Framework Evaluation using DESMET Feature Analysis

Authors: Asyraf Wahi Anuar, Nazri Kama, Azri Azmi, Hazlifah Mohd Rusli, Yazriwati Yahya

PAGE 437 – 452

Paper 53: A Novel Readability Complexity Score for Gujarati Idiomatic Text

Authors: Jatin C. Modh, Jatinderkumar R. Saini, Ketan Kotecha

PAGE 453 – 459

Paper 54: Importance of Memory Management Layer in Big Data Architecture

Authors: Maha Dessokey, Sherif M. Saif, Hesham Eldeeb, Sameh Salem, Elsayed Saad

PAGE 460 – 466

Paper 55: Structural Equation Modelling for Validating Disruptive Factors in Livestock Supply Chain

Authors: Nur Amlia Abd Majid, Noraidah Sahari, Nur Fazidah Elias, Hazura Mohamed, Latifah Abd Latib, Khairul Firdaus Ne'matullah

PAGE 467 – 476

Paper 56: IoT Enabled Smart Parking System for Improvising the Prediction Availability of the Parking Space

Authors: Anchal, Pooja Mittal

PAGE 477 – 486

Paper 57: An Optimized Kernel MSVM Machine Learning-based Model for Churn Analysis

Authors: Pankaj Hooda, Pooja Mittal

PAGE 487 – 494

Paper 58: Study on Feature Engineering and Ensemble Learning for Student Academic Performance Prediction

Authors: Du Xiaoming, Chen Ying, Zhang Xiaofang, Guo Yu

PAGE 495 – 502

Paper 59: Development of Hausa Acoustic Model for Speech Recognition

Authors: Umar Adam Ibrahim, Moussa Mahamat Boukar, Muhammad Aliyu Suleiman

PAGE 503 – 508

Paper 60: New Method for 4D Reconstruction of Medical Images

Authors: Lamyae MIARA, Said BENOMAR ELMDEGHRI, Mohammed Oucamah CHERKAOUI MALKI

PAGE 509 – 518

Paper 61: An Adaptive Approach for Preserving Privacy in Context Aware Applications for Smartphones in Cloud Computing Platform

Authors: H. Manoj T. Gadiyar, Thyagaraju G. S, R. H. Goudar

PAGE 519 – 529

Paper 62: Digital Learning Tools for Security Inductions in Mining Interns: A PLS-SEM Analysis

Authors: Jose Julian Rodriguez-Delgado, Patricia Lopez-Casaperalta, Mario Gustavo Berrios-Espezua, Alejandro Marcelo Acosta-Quelopana, Jose Sulla-Torres

PAGE 530 – 536

Paper 63: Enhanced Symbol Recognition based on Advanced Data Augmentation for Engineering Diagrams

Authors: Ong Kai Bin, Yew Kwang Hooi, Said Jadid Abdul Kadir, Haruhiro Fujita, Luqman Hakim Rosli

PAGE 537 – 546

Paper 64: A Graph-oriented Framework for Online Analytical Processing

Authors: Abdelhak KHALIL, Mustapha BELAISSAOUI

PAGE 547 – 555

Paper 65: Sentiment Analysis to Explore User Perception of Teleworking in Saudi Arabia

Authors: Malak Nazal Alotaibi, Zahyah H. Alharbi

PAGE 556 – 563

Paper 66: Framework for Development of 3D Temple Objects based on Photogrammetry Method

Authors: Herman Tolle, Ratih Kartika Dewi, Komang Candra Brata, Benyamin Perdamean

PAGE 564 – 571

Paper 67: Dual-fast Greedy Heuristic Algorithm for Green ICT

Authors: Inas Abuqaddom

PAGE 572 – 576

Paper 68: Parallel Improved Genetic Algorithm for the Quadratic Assignment Problem

Authors: Huda Alfaifi, Yassine Daadaa

PAGE 577 – 583

Paper 69: Modeling for Car Quality Complaint Classification based on Machine Learning

Authors: Chen Xiao Yu, Hou Xia, Zhang Xiao Min, Song Ying

PAGE 584 – 591

Paper 70: Identifying Influential Nodes with Centrality Indices Combinations using Symbolic Regressions

Authors: Mohd Fariduddin Mukhtar, Zuraida Abal Abas, Amir Hamzah Abdul Rasib, Siti Haryanti Hairol Anuar, Nurul Hafizah Mohd Zaki, Ahmad Fadzli Nizam Abdul Rahman, Zaheera Zainal Abidin, Abdul Samad Shibghatullah

PAGE 592 – 599

Paper 71: Improving Computational Thinking in Nursing Students through Learning Computer Programming

Authors: Leticia Laura-Ochoa, Norka Bedregal-Alpaca, Elizabeth Vidal

PAGE 600 – 605

Paper 72: Improving Social Engineering Awareness, Training and Education (SEATE) using a Behavioral Change Model

Authors: Azaabi Cletus, Benjamin Weyory, Alex Opoku

PAGE 606 – 613

Paper 73: Evaluating Learning Management System based on PACMAD Usability Model: Brighten Mobile Application

Authors: Masyura Ahmad Faudzi, Zaihisma Che Cob, Ridha Omar, Sharul Azim Sharudin

PAGE 614 – 621

Paper 74: Voice Biometrics for Indonesian Language Users using Algorithm of Deep Learning CNN Residual and Hybrid of DWT-MFCC Extraction Features

Authors: Haris Isyanto, Ajib Setyo Arifin, Muhammad Suryanegara

PAGE 622 – 634

Paper 75: Hybrid Deformable Convolutional with Recurrent Neural Network for Optimal Traffic Congestion Prediction: A Fuzzy Logic Congestion Index System

Authors: Sara Berrouk, Abdelaziz El Fazziki, Mohammed Sadgal

PAGE 635 – 651

Paper 76: Impact of the Pandemic on the Development and Regulation of Electronic Commerce in Russia

Authors: Svetlana Panasenko, Maisa Seifullaeva, Ibragim Ramazanov, Elena Mayorova, Alexander Nikishin, A. M. Vovk

PAGE 652 – 658

Paper 77: Application for a Waste Management via the QR-Code System

Authors: Pichit Wandee, Zakon Bussabong, Seksit Duangkum

PAGE 659 – 664

Paper 78: Empirical Study of a Spatial Analysis for Prone Road Traffic Accident Classification based on MCDM Method

Authors: Anik Vega Vitianingsih, Zahriah Othman, Safiza Suhana Kamal Baharin, Aji Suraji

PAGE 665 – 679

Paper 79: Application of the Fuzzy Delphi Method to Identify and Prioritize the Social-Health Family Disintegration Indicators in Yemen

Authors: Abed Saif Ahmed Alghawli, Abdualmajed A. Al-khulaidi, Adel A. Nasser, Nesmah A. AL-Khulaidi, Faisal A. Abass

PAGE 680 – 691

Paper 80: A Penetration Testing on Malaysia Popular e-Wallets and m-Banking Apps

Authors: Md Arif Hassan, Zarina Shukur, Masnizah Mohd

PAGE 692 – 703

Paper 81: Demand Forecasting Model using Deep Learning Methods for Supply Chain Management 4.0

Authors: Loubna Terrada, Mohamed El Khaili, Hassan Ouajji

PAGE 704 – 711

Paper 82: Improved Deep Learning Performance for Real-Time Traffic Sign Detection and Recognition Applicable to Intelligent Transportation Systems

Authors: Anass BARODI, Abderrahim Bajit, Abdelkarim ZEMMOURI, Mohammed Benbrahim, Ahmed Tamtaoui

PAGE 712 – 723

Paper 83: Research on Students' Course Selection Preference based on Collaborative Filtering Algorithm

Authors: Mustafa Man, Jianhui Xu, Ily Amalina Ahmad Sabri, Jiaxin Li

PAGE 724 – 733

Paper 84: Intelligent Interfaces for Assisting Blind People using Object Recognition Methods

Authors: Jamil Abedalrahim Jamil Alsyayadeh, Irianto, Maslan Zainon, Hasvinii Baskaran, Safarudin Gazali Herawan

PAGE 734 – 741

Paper 85: Application of Random Forest Regression with Hyper-parameters Tuning to Estimate Reference Evapotranspiration

Authors: Satendra Kumar Jain, Anil Kumar Gupta

PAGE 742 – 750

Paper 86: Abnormal Event Detection using Additive Summarization Model for Intelligent Transportation Systems

Authors: G. Balamurugan, J. Jayabharathy

PAGE 751 – 757

Paper 87: Relational Deep Learning Detection with Multi-Sequence Representation for Insider Threats

Authors: Abdullah Alshehri

PAGE 758 – 765

Paper 88: An Improved Label Initialization based Label Propagation Method for Detecting Graph Clusters in Complex Networks

Authors: Jyothimon Chandran, V Madhu Viswanatham

PAGE 766 – 776

Paper 89: Natural Language Processing for the Analysis Sentiment using a LSTM Model

Authors: Achraf BERRAJAA

PAGE 777 – 785

Paper 90: Alarm System using Image Processing to Prevent a Patient with Nasogastric Tube Feeding from Removing Tube

Authors: Amonrat Prasitsupparote, Pakorn Pasitsuparoad

PAGE 786 – 792

Paper 91: Multileveled ALPR using Block-Binary-Pixel-Sum Descriptor and Linear SVC

Authors: B. Lavanya, G. Lalitha

PAGE 793 – 801

Paper 92: Genetic Algorithms Applied to the Searching of the Optimal Path in Image-based Robotic Navigation Environments

Authors: Fernando Martinez Santa, Fredy H. Martinez Sarmiento, Holman Montiel Ariza

PAGE 802 – 808

Paper 93: Anomaly Detection using Network Metadata

Authors: Khaled Mutmbak, Sultan Alotaibi, Khalid Alharbi, Umar Albalawi, Osama Younes

PAGE 809 – 814

Paper 94: Correcting Arabic Soft Spelling Mistakes using BiLSTM-based Machine Learning

Authors: Gheith Abandah, Ashraf Suyyagh, Mohammed Z. Khedher

PAGE 815 – 829

Paper 95: Prediction of Presence of Brain Tumor Utilizing Some State-of-the-Art Machine Learning Approaches

Authors: Mitrabinda Khuntia, Prabhat Kumar Sahu, Swagatika Devi

PAGE 830 – 840

Paper 96: Mining Hidden Partitions of Voice Utterances using Fuzzy Clustering for Generalized Voice Spoofing Countermeasures

Authors: Sarah Mohammed Altuwayjiri, Ouiem Bchir, Mohamed Maher Ben Ismail

PAGE 841 – 849

Paper 97: PhishRepo: A Seamless Collection of Phishing Data to Fill a Research Gap in the Phishing Domain

Authors: Subhash Ariyadasa, Shantha Fernando, Subha Fernando

PAGE 850 – 865

Paper 98: A Novel Code Completion Strategy

Authors: Hayatou Oumarou, Ousmanou Dahirou

PAGE 866 – 871

Paper 99: Revisiting Polyglot Persistence: From Principles to Practice

Authors: Omar Lajam, Salahadin Mohammed

PAGE 872 – 882

Paper 100: Computer Vision: The Effectiveness of Deep Learning for Emotion Detection in Marketing Campaigns

Authors: Shaldon Wade Naidoo, Nalindren Naicker, Sulaiman Saleem Patel, Prinavin Govender

PAGE 883 – 890

Paper 101: Using Machine Learning Techniques to Predict Bugs in Classes: An Empirical Study

Authors: Musaad Alzahrani

PAGE 891 – 897

Paper 102: Transformer-based Models for Arabic Online Handwriting Recognition

Authors: Fakhraddin Alwajih, Eman Badr, Sherif Abdou

PAGE 898 – 905

Paper 103: Detection of Android Malware App through Feature Extraction and Classification of Android Image

Authors: Mohd Abdul Rahim Khan, Nand Kumar, R C Tripathi

PAGE 906 – 914

Paper 104: A Hybrid Heuristic for a Two-Agent Multi-Skill Resource-Constrained Scheduling Problem

Authors: Meya Haroune, Cheikh Dhib, Emmanuel Neron, Ameer Soukhal, Hafed Mohamed Babou, Farouk Mohamedade Nanne

PAGE 915 – 928

Paper 105: Transformer based Model for Coherence Evaluation of Scientific Abstracts: Second Fine-tuned BERT

Authors: Anyelo-Carlos Gutierrez-Choque, Vivian Medina-Mamani, Eveling Castro-Gutierrez, Rosa Nunez-Pacheco, Ignacio Aguaded

PAGE 929 – 937

Paper 106: Modeling and Simulation of Adaptive Traffic Control System for Multi-Intersection Management using Cellular Automaton and Queuing System

Authors: Salma EL BAKKAL, Abdallah LAKHOULI, El Hassan ESSOUFI

PAGE 938 – 947

Paper 107: Non-Parametric Stochastic Autoencoder Model for Anomaly Detection

Authors: Raphael Alampay, Patricia Angela Abu

PAGE 948 – 959

Paper 108: COVID-19 Cases Detection from Chest X-Ray Images using CNN based Deep Learning Model

Authors: Md Amirul Islam, Giovanni Stea, Sultan Mahmud, Kh. Mustafizur Rahman

PAGE 960 – 971

Paper 109: BERT-based Approach to Arabic Hate Speech and Offensive Language Detection in Twitter: Exploiting Emojis and Sentiment Analysis

Authors: Maha Jarallah Althobaiti

PAGE 972 – 980

Paper 110: A Survey of Sink Mobility Models to Avoid the Energy-Hole Problem in Wireless Sensor Networks

Authors: Ghada Al-Mamari, Fatma Bouabdallah, Asma Cherif

PAGE 981 – 993

Paper 111: End-to-End Car Make and Model Classification using Compound Scaling and Transfer Learning

Authors: Omar BOURJA, Abdelilah MAACH, Zineb ZANNOUTI, Hatim DERROUZ, Hamza MEKHZOUM, Hamd AIT ABDELALI, Rachid OULAD HAJ THAMI, Francois BOURZEIX

PAGE 994 – 1001

Paper 112: Cache Complexity of Cache-Oblivious Approaches: A Review and Extension

Authors: Inas Abuqaddom, Sami Serhan, Basel A. Mahafzah

PAGE 1002 – 1009

Paper 113: OvSbChain: An Enhanced Snowball Chain Approach for Detecting Overlapping Communities in Social Graphs

Authors: Jayati Gulati, Muhammad Abulaish, Sajid Yousuf Bhat

PAGE 1010 – 1019

Paper 114: Improving the Computational Complexity of the COOL Screening Tool

Authors: Mohamed Ghalwash

PAGE 1020 – 1027

Paper 115: A Lightweight Verifiable Secret Sharing in Internet of Things

Authors: Likang Lu, Jianzhu Lu

PAGE 1028 – 1035

Implementation of Data Mining on a Secure Cloud Computing over a Web API using Supervised Machine Learning Algorithm

Data Mining in a Secure Cloud Computing Environment through Restful API

Tosin Ige¹

Department of Computer Science
University of Texas at El Paso
Texas, USA

Sikiru Adewale²

Department of Computer Science
Virginia Technological University
Virginia, USA

Abstract—Ever since the era of internet had ushered in cloud computing, there had been increase in the demand for the unlimited data available through cloud computing for data analysis, pattern recognition and technology advancement. With this also bring the problem of scalability, efficiency and security threat. This research paper focuses on how data can be dynamically mine in real time for pattern detection in a secure cloud computing environment using combination of decision tree algorithm and Random Forest over a restful Application Programming Interface (API). We are able to successfully implement data mining on cloud computing bypassing or avoiding direct interaction with data warehouse and without any terminal involve by using combination of IBM Cloud storage facility, Amazon Web Service, Application Programming Interface and Window service along with a decision tree and Random Forest algorithm for our classifier. We were able to successfully bypass direct connection with the data warehouse and cloud terminal with 94% accuracy in our model.

Keywords—Cloud computing; data warehouse; data mining; window service; Web API; machine learning algorithm; secure cloud computing

I. INTRODUCTION

As we all know that Knowledge discovery is the nontrivial extraction of implicit, previously unknown, and potentially useful information from data [1]. There is no doubt that the availability of billions of data in the cloud had open floodgates of opportunity in which model can be trained and learn by itself over time thereby enable machines and intelligent elements to make crucial and important decision without human intervention. As millions of data are constantly being stored and retrieve through cloud computing on daily bases. There is no doubt that data mining delivers a powerful competitive advantage [6] for any company or industry. With these also arises the challenge on how to make use of the unlimited data available through data mining on cloud computing, and how it can be done in a highly secure, scalable and efficient manner to combat several security threats associated with cloud computing.

In modern research and intelligent systems, the important of data can never be overemphasized across wide range of industries, especially when we consider the fact that the

traditional E-commerce businesses and industry were influenced by cloud computing in technical architecture, service modes and the industrial chain [4], economist needs to know how they can use readily available data in the cloud to predict consumers need and behavior, meteorologists need those data to make future weather forecast and predict or detect climatic change, government agency needs those data to make effective policy, police and intelligence agency needs data for background check and so on. All government and non-government parastatal depends on data from the cloud for one reason or the other. This further brings about the importance and urgency of mining data to detect previously unknown pattern in a scalable and efficient manner through cloud computing to address daily needs with the guarantee of maximum security, authorization and authentication which can be affected through a restful web API.

In this research, the use of constantly updating real time data and without any connection to the cloud data warehouse for maximal data protection and dynamic predictive pattern with high accuracy were implemented. To achieve this objective, an application Programming Interface (API) and Background window service was developed to detect and fetch new and updated records from data warehouse, transformed to json and written to a file on the cloud. This ensures maximum security to the data as there is no direct connection to the data warehouse where data are constantly being pulled from. We used IBM Cloud object storage facility to host our csv file while the data warehouse is also on the cloud but from another channel entirely. Then a restful web API service was developed using Django python framework which was deployed to Amazon Web Services (AWS). Since the newly created API will be constantly pulling data from the data warehouse and writing directly to into the CSV file which is on the IBM Cloud storage facility. The service can easily be overwhelmed, to prevent the service from being overwhelm, we developed a background service on a Microsoft console using Microsoft .NET Technology programming language.

Any record the web API misses or delayed in picking from the warehouse, the background service will pick it up. Hence, they complement each other and ensure efficiency so that the web API service is not overwhelmed. With this in place, there

is a constantly updated and up to date data on the CSV file hosted in the IBM cloud storage facility which is now serves as the primary source of data. This ensure accessibility to updated records at every stage of our data mining activities, as there is an existing restful API and background window service that fetches data from the warehouse to the hosted file in cloud environment from which data is being pulled from.

Combination of decision tree algorithm and Random Forest for classifier was implemented coupled with a graphic user interface (GUI) that automatically processed and displays newly discovered patterns in a graphical format on the screen at each and every update on new data.

There are four main objectives here which are:

- 1) Bypassing direct connection and retrieval from the data warehouse through a middle ware.
- 2) Real time access to dynamic data in an enabling cloud environment.
- 3) We want to shield the warehouse for maximum security and while also avoiding data lock.
- 4) Automation of the detection of any new pattern from the dataset and projection to the screen with graphical illustration and analysis.

To ensure successful accomplishment of the fourth objective, a scheduler called Python scheduling library was used so that the whole process is automatically initiated and repeated immediately a new pattern is detected.

II. BACKGROUND STUDY

The importance of cloud computing cannot be overemphasize as it includes; unlimited storage, provisioning and updating, guaranteed privacy more security [13]. Also, it is possible for users of cloud services to optimize server utilization, dynamic scalability, and minimize the development of new application life cycles [14]. In addition to the numerous benefits in cloud computing, there are also problems as a result of cloud outage since data storage is centralize in the cloud which can paralyze a company business [15], also attack on the integrated cloud environment can cause loss of data and finance for both the service providers and subscribers. There are also other risks associated with computing on cloud environment which includes the issues of threats to data security information confidentiality [5], and the possibility of information leakage and vulnerability [12].

In today's data mining, multiple data streams are generally complex than single data streams [2] considering the security architecture and cloud computing environment, the complexity increases when we consider the cost and benefit, reliability, cloud migration and inter clouding [3], While compute/storage scaling, data parallelism, virtualization, MapReduce, RIA, SaaS and Mashups are likewise also important in data mining [7] not all are always implemented. Although there had been improvement in the technology and services of the major Cloud Computing Platforms likes amazon Relational Database Service, Amazon Simple Queue Service, Amazon SimpleDB, Amazon Web Services, AppScale, Azure Services Platform, Caspio, CloudControl, Cordys Process Factory, Engine Yard, Force.com, FreedomBox, Google App Engine, Heroku, Hybrid

Web Cluster, OrangeScape, Platform as a service, Rackspace Cloud, Rollbase, Squarespace, Sun Cloud, Vertebra (cloud computing framework), Wolf Frameworks [8], some of the persistent problems associated with data mining on cloud computing are as a result of the existing method adopted in the industry and scientific world at large. Current cloud computation for data mining needed service provider to provide interface for user. The user does not need to bother about the infrastructure. It enables the retrieval of useful previously unknown data from integrated data warehouse, in such a way that users doesn't need to border about infrastructure, storage, or configuration and maintenance, the provider handles that. It is based on retrieving directly from the integrated data warehouse.

[5] discusses the importance of large item sets within a cluster and its implication for effectiveness, it doesn't address the several issues and vulnerabilities in cloud computing while [9] in his research work "Data mining in Cloud Computing", relies on extraction of previously unknown or meaningful pattern from unstructured or semi-structured data from the web sources; "The analysis steps in the Knowledge Discovery and Databases process" [11]. It listed three stages of research involving data warehouse which are staging, integration, and accessibility for the purpose of reporting and analysis in the Review of Data Mining Techniques in Cloud Computing Database by [10].

In addition to the problems associated with data mining on cloud computing in which majority of the problems are due to existing methods of data mining on cloud computing. The existing methods also creates a wide gap between data mining on cloud computing and application Programming Interface (API) which are yet to be closed. In the existing method we are yet to see an instance in which we call we only need to call an API endpoint, and then have everything done for us.

III. RESEARCH METHODOLOGY

Data used on this research work was obtained from social network on GitHub. To begin this research, four basic things were paramount;

- 1) Subscription to IBM Cloud Object Storage facility to host our CSV file.
- 2) Setting up a data warehouse on cloud from another channel different from IBM.
- 3) Development of a background window service using .NET Technology.
- 4) Implementation of a web service to be consumed in the cloud using python.
- 5) Scheduler using python scheduler library to trigger the web API at interval.

In order to prevent the web service from being overwhelm due to multiple calling of the endpoint at regular interval, we developed a background window service to support the restful API service. Both the web API and the window service are picking records from the integrated data warehouse and pushing to the CSV file on the IBM Object Cloud Storage facility. They automatically pick new records to the CSV, and if a record is modified, it will be picked and modified on the

CSV as well. The essence of the scheduler which was written in python is to be calling the web API at regular interval to check and push from the integrated data warehouse to the CSV file.

With the successful setting up of cloud environment and the necessary software programs being in fully execution, we proceeded by adopting the following data cleansing and preparation techniques;

Data cleaning: We use preprocessing and cleaning methods to remove incomplete data that might cause system failure and also affect output prediction. Rows containing missing values where completely removed. We also use different methods to identify and remove noisy data, outliers, and other factors which can influence the output result.

Data Reduction for Data Quality: In order to maintain data integrity, we needed to deal with all rows containing null value; hence we opted for python library tool called PyCaret. We have two options which are either to automatically fill all the null values or to remove any row(s) containing null values weighted. Having weighted the risk involves in both, we decided to remove any row with null or empty value from the data, and this was done using PyCaret python library.

Data Transformation: For us to make our data to acceptable format for easy data mining and pattern recognition, it needed to undergo some data transformation. To ensure data is fully transformed to acceptable format, we used Discretization, normalization, and data aggregation technique.

Data Mining: Having successfully set up our apparatus which includes IBM cloud object storage facility, running background window service, active web service, scheduler, and with the data being thoroughly pre-processed and transformed.

Unlike current data mining in cloud computing process in which, one needs to make direct connect to the data warehouse or directly call the csv file. We only called our restful API endpoint (Fig. 1) which was developed and hosted on the cloud.

This ensures a higher level of security and control over the data, the only thing that needed to be called is the endpoint of our API which automatically displays the data as seen in Fig. 1 displaying the first five (5) records in the data using python library in panda. (Fig. 2) shows the text representation of the selected features using decision tree algorithm.

```
In [3]: data = pd.read_csv('http://cs4430.meritgotech.com/customers/social')
data.head()

Out[3]:
```

	User ID	Gender	Age	EstimatedSalary	Purchased
0	15624510	Male	19	19000	0
1	15810944	Male	35	20000	0
2	15686575	Female	26	43000	0
3	15603246	Female	27	57000	0
4	15804002	Male	19	76000	0

Fig. 1. Preview of First Five Rows of Dataset.

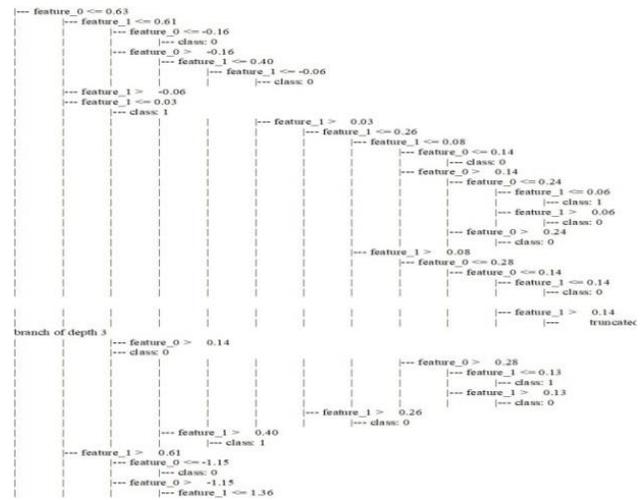


Fig. 2. Text Representation of the Features using Decision Tree.

Optimized Decision tree algorithm was used to train the model having slatted the data into two equal part, half of it to train the model and the remaining half for testing of the model as seen in (Fig. 3).

```
feature_cols = ['Age','EstimatedSalary']
X = data.iloc[:,[2,3]].values y = data.iloc[:,4].values
#split the dataset into training and test
X_train, X_test, y_train, y_test = train_test_split(X,y,test_size = 0.25, random_state= 0)
#perform feature scaling
from sklearn.preprocessing import StandardScaler sc_X = StandardScaler()
X_train = sc_X.fit_transform(X_train) X_test = sc_X.transform(X_test)
#fit the model in the decision tree classifier from sklearn.tree import DecisionTreeClassifier classifier = DecisionTreeClassifier()
classifier = classifier.fit(X_train,y_train)
```

Fig. 3. Code Snippet for Important Feature Selection and Training of Model.

We are able to obtain 90% accuracy on testing our model, after which we decide to optimize for more accuracy and better performance as seen in Fig. 4.



Fig. 4. Model Accuracy and Test Performance Evaluation.

We are able to obtain accuracy of 94% after which we decided to visualize the performance of the model.

We needed to know the level of fitting, because having an accuracy of 94% can be as a result of over fitting of the model with the training set of data, hence we decided to do two additional things. Firstly, we implemented Random Forest Algorithm for training the data again to check performance and then optimize, it gives an accuracy of 92%. Secondly, we fed another set of data in the format of the trained data but new to the model, the model performed very well with accuracy. So, we proceeded to measure the performance of the model using confusion matrix, classification report, and accuracy score which gives good indication of optimal performance. We are in

a dilemma either to ensure about 100% accuracy or over fitting because in supervise machine learning, high accuracy of almost 100% can be as a result of over fitting of the model which we want to avoid by possible means. So, since we are able to feed the model with new set of data which had not been previously fed to the model for which it performed very well with high rate of accuracy. So, we gave priority to avoid over fitting of the model than the accuracy of the model since high accuracy for supervised learning can be as a result of over fitting of the model for which the model becomes less accurate or behaved weird when fed with unfamiliar data. The validation report can be seen in Fig. 5.

```
-----
Mean Absolute Error: 0.09166666666666666
-----
Mean Squared Error: 0.09166666666666666
-----
Root Mean Squared Error: 0.30276503540974914
-----

[[61 4]
 [ 7 48]]
-----
      precision    recall  f1-score   support

0         0.90      0.94      0.92         65
1         0.92      0.87      0.90         55

 accuracy
macro avg      0.91      0.91      0.91         120
weighted avg   0.91      0.91      0.91         120
```

Fig. 5. Validation, Cross Validation and the Estimation of Mean Square Error (MSE).

IV. CONCLUSION

In this applied research, we are able to achieve four goals:

- 1) Avoid direct interaction with integrated data warehouse.
- 2) We are able to access data in a secure cloud computing environment.
- 3) We are able to close the existing gap in data mining and web API, as all we needed to call is the endpoint of our web API. No need of terminal, or uniform resource locator (URL) of any csv is involved. It is simply over a web API.
- 4) As the patterns changes in the background, the algorithm automatically adjusts with accuracy of ninety-four (94) percent after optimization.

We are able to implement data mining in a secure cloud computing environment with accuracy of ninety-four (94) percent after optimization in our decision tree algorithm over web API. All that an AI engineer, data scientist, or machine learning engineer needs is just the endpoint of the API. This removes complexity while at the same time simplifying the whole process, it also add additional layer of security, and also remove unnecessary bottleneck as scientist and engineers will be able to concentrate more on optimizing their algorithm and model for optimal result since only API endpoint is what is needed to be called.

We hope that this will be de-facto standard in the data mining, machine learning, data science and other similar industry at large.

V. LIMITATION AND FUTURE RESEARCH WORK

This research work is based on the quantity of data available to us. Also, as Infrastructure As a Service Provider (IAAS), Software As a Service Provider (SAAS), and Platform As a Service Provider (PAAS) continually improves their service for more secured and enhanced cloud computing environment, over the times, this can affect the performance of the model and the overall architecture, also as more data becomes available, there is possibility of more false positive and this directly impacts the bias-variance tradeoff.

So, there is future research work of developing what we called intelligent model, model which can detect availability or changes in data, detect changes in the cloud computing environment and then re-trained and re-adjust itself over the time in line with those changes.

REFERENCES

- [1] W. J. Frawley, G. Piatetsky-shapiro, and C. J. Matheus. Knowledge discovery in databases: an overview, 1992.
- [2] W.Wu and L. Gruenwald. Research issues in mining multiple data streams. In Proceedings of the First International Workshop on Novel Data Stream Pattern Mining Techniques, Stream KDD'10, pages 56–60, New York, NY, USA, 2010.ACM.
- [3] David E.Y. Sarna, Implementing And Developing Cloud Computing Applications, CRC Press <https://cwiki.apache.org/MAHOUT/kmeans-clustering.html>.
- [4] Weiss, A. (2007). Computing in the Clouds Networker.
- [5] Wang, K., Xu, C. & Liu, B. (1999). Clustering transactions using large items, in, CIKM '99: Proceedings of the Eighth International Conference on Information and Knowledge Management", ACM Press, New York, NY, USA, pp. 483–490.
- [6] Berson, Alex, Stephen Smith, and Kurt Thearling. Building data mining applications for CRM. New York: McGraw-Hill, 2000.
- [7] Moving To The Cloud: Developing Apps in the new world of cloud computing , By Dinkar Sitaram, Geetha Manjunath.
- [8] The Cloud Computing Handbook Everything You Need to Know about Cloud Computing, By Todd Arias.
- [9] Data mining in cloud computing by Ruxandra-Ştefania PETRE, Link: https://www.dbjournal.ro/archive/9/9_7.pdf.
- [10] Review of Data Mining Techniques in Cloud Computing Database, by Astha Pareek1, Manish Gupta2.
- [11] M. M. Nodine, A. H. H. Ngu, A. Cassandra and W. G. Bohrer, "Scalable semantic brokering over dynamic heterogeneous data sources in InfoSleuth/spl trade/," in IEEE Transactions on Knowledge and Data Engineering, vol. 15, no. 5, pp. 1082-1098, Sept.-Oct. 2003, doi: 10.1109/TKDE.2003.1232266.
- [12] Andrei Tchernykh, Uwe Schwiiegelsohn, El-ghazali Talbi, Mikhail Babenko, Towards understanding uncertainty in cloud computing with risks of confidentiality, integrity, and availability, Journal of Computational Science, Volume 36, 2019, 100581, ISSN 1877-7503, <https://doi.org/10.1016/j.jocs.2016.11.011>.
- [13] Ko, H., Hofer, S., Pichler, B. et al. Functional specificity of local synaptic connections in neocortical networks. Nature 473, 87–91 (2011). <https://doi.org/10.1038/nature09880>.
- [14] Al-Ruithe, Majid & Benkhelifa, Elhadj & Hameed, Khawar. (2019). A systematic literature review of data governance and cloud data governance. Personal and Ubiquitous Computing. 23. 10.1007/s00779-017-1104-3.
- [15] Godhankar, P.B. and Gupta, D. (2014) Review of Cloud Storage Security and Cloud Computing Challenges. International Journal of Computer Science & Information Technology, 5, 528-533.

AI Powered Anti-Cyber Bullying System using Machine Learning Algorithm of Multinomial Naïve Bayes and Optimized Linear Support Vector Machine

Interception of Cyberbully Contents in a Messaging System by Machine Learning Algorithm

Tosin Ige¹

Department of Computer Science
University of Texas at El Paso, Texas, USA

Sikiru Adewale²

Department of Computer Science
Virginia Technological University, Virginia, USA

Abstract—“Unless and until our society recognizes cyber bullying for what it is, the suffering of thousands of silent victims will continue.” ~ Anna Maria Chavez. There had been series of research on cyber bullying which are unable to provide reliable solution to cyber bullying. In this research work, we were able to provide a permanent solution to this by developing a model capable of detecting and intercepting bullying incoming and outgoing messages with 92% accuracy. We also developed a chatbot automation messaging system to test our model leading to the development of Artificial Intelligence powered anti-cyber bullying system using machine learning algorithm of Multinomial Naïve Bayes (MNB) and optimized linear Support Vector Machine (SVM). Our model is able to detect and intercept bullying outgoing and incoming bullying messages and take immediate action.

Keywords—Cyberbullying; anti cyberbullying; machine learning; NLP; social media; multinomial Naïve Bayes; support vector machine

I. INTRODUCTION

Hatred, violence, and hostility in modern world can take several form [4],[5],[6],[2], one of which is cyber bullying using modern day technology medium. While the era of internet had brought in tremendous innovation and improvements to our daily activities and overall way of life, it had also opened floodgates for cyber bullying. Any devastating act in the mode of aggressive or abusive behavior toward people regarding digital interactions is cyber bullying [1] Fig. 1

The impact of social media like Instagram, Facebook, Twitter, WhatsApp, etc. on daily basis cannot be over emphasize as they had greatly influence modern way of communication As useful as social media is, it is a medium for promoting hatred, harassment, racism, etc. which is currently affecting millions of people across the globe.

Statistical record from 2019 Cyber bullying Data shows that 95% of teens in the U.S. are online, and the vast majority has access to internet on their mobile device, makes social media platform the most common medium for cyber bullying [11]. About 37% of young people between the ages of 12 and 17 have been bullied online. 30% have had it happen more than once [8],[9],[10]. The impact of cyber bullying is very visible in the world today as it had result in several hatred, trauma,

depression, and untimely death. Ryan Halligan (1989–2003), age 13, was an American student from Essex Junction, Vermont, who died by suicide at the age of 13 after allegedly being bullied by his classmates in person and online. While Jeff Weise (1988–2005), age 16, who was also an American high school student who committed the Red Lake shootings killing nine people and himself by suicide after being severally attack by cyber bullying Fig. 2.

There had been several measures and implementations put in place to prevent cyber bullying as a solution but none of them have actually solve cyber bully as the effect of cyber bullying is still obvious in our society.

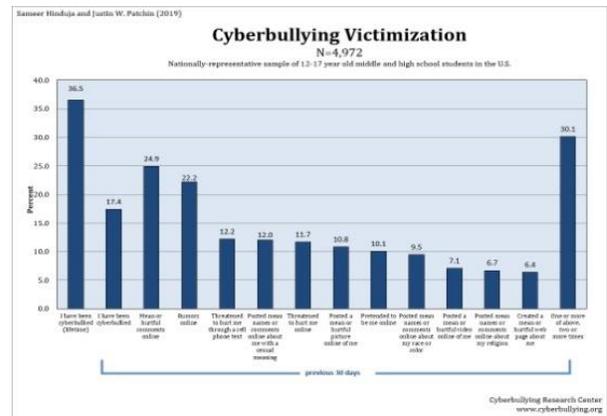


Fig. 1. Statistical Analysis and Victim of Cyberbullying.

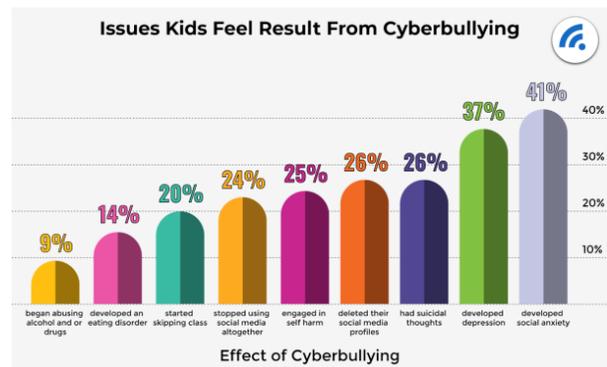


Fig. 2. Statistical Effects of Cyberbullying.

In this research work, Multinomial Naïve Bayes and Optimized Linear Support Vector Machine of machine learning algorithm was successfully used to implement an Artificial Intelligent Powered anti-cyber bullying system capable of detecting and intercepting online cyberbullying messages and filtered [3] them so that the intending recipient doesn't receive it. To ensure a better coverage across billions of device worldwide, we deployed our model and exposed it to a web application programming interface (API). The REST API was developed with Flask python framework, so the API load our deployed model, call it with parameterized input, and then indicated the status if it is cyberbullying or not using natural language processing capability of machine learning for detection and interception. All that any social media website like Facebook, Twitter, Instagram, LinkedIn, Snapchat, YouTube, WhatsApp, etc. needed to do is to integrate it on their platform by calling the API alone. With this implementation, the problem of cyber bullying had been laid to rest. In order to test our implementation in real time, we developed a chatbot automation messaging system, and then use it to consume our restful API service, we got an accuracy of 92% and the model was able to automatically detect and intercept any outgoing and incoming bullying message and then ensure that the intending receiver doesn't receive the message. We also estimate the time taken by the restful API service to detect, intercept, and process messages and discovered that it is negligible fraction of seconds which simply that our implementation can be implemented, deployed, and use in real time.

II. BACKGROUND STUDY

Over the years, there had been several efforts to address the issue of cyber bullying, but none had actually enforce the prevention which is why the rate of cyberbully had remain so high in our society. The authors in [14] from the MIT Media Lab, led by KarthikDinakar implemented an algorithm based on clustering and classification that is able to detect and categorize group of words in an online interaction [3]. The algorithm correctly categorizes contents in online interaction with high accuracy. It could not be used to combat cyber bullying both in real time. Several models and algorithm had been developed from various research labs, but none had been able to combat cyber bullying in real time. Logical probabilistic model of approach was used to develop a socio-linguistic model capable of detecting cybeybullying and the role play in the context of the conversation [15], Natural Language processing method of approach had been employed to identify cyber bullying on chat conversation [11], while Raisi uses Co-trained ensembles for weakly supervised bullying detection of embedded models used RNN and node2vec learners for detecting harassment and bullying from text base data [7].

As different research had been carried out from different research labs and centers, many of which involves the use of Different machine and deep learning techniques to catch word phrases and slangs like k Nearest Neighbor (kNN), Linear Regression (LR), Random Forests (RF), Logistic Regression (LogR), Boosting (Bos), Bagging (Bgg), Adaboost (ADB), Multiple Regression (MR), Maximum Entropy (MaxE), etc.

are being used for detection of cyberbullying on social media [12],[13].

Kainat et al, in 2021 proposed machine learning based algorithm which detects harassment actively and alert user to take action against it [13] But Kainat et al proposal is not feasible in real time and does not solve the problem of cyberbully due to the following reasons:

- 1) It does not intercept the incoming message: A real time anti cyberbully must intercept the incoming message before getting to the receiver.
- 2) The receiver have to take action to delete the content: If someone harasses me online, even though I delete the message, the fact that I already receive it will have psychological effect on me.
- 3) Reliance on csvfile from both sender and receiver which is not feasible due to cost and scalability.
- 4) No API exposure for wide coverage.

In this research work, we were able to implement anti-cyberbully system, which is able to automatically intercept an incoming message before getting to the receiver and take necessary action. Three major steps were involved:

- 1) We used Multinomial Naïve Bayes (MNB) and Optimized Linear Support vector Machine (svm) to train our model.
- 2) Deployed the model.
- 3) Build a Restful API service using Flask framework of Python.

III. RESEARCH METHODOLOGY

For this research, we used different sources of dataset collection which are related to cyber-bullying. The data is from different social media platforms like Kaggle, Twitter, Wikipedia Talk pages and YouTube. The data contain text and labeled as bullying or not. The data contains different types of cyber-bullying like racism, hate speech, aggression, insults and toxicity. Having set up our cloud environment and the necessary programs written, we proceeded by adopting the standard machine learning data cleansing and preparation techniques Fig. 3.

Data cleaning: We use preprocessing and cleaning methods to remove incomplete data that might cause system failure and also affect output prediction. Rows containing missing values where completely removed. We also use different methods to identify and remove noisy data, outliers, and other factors which can influence the output result.

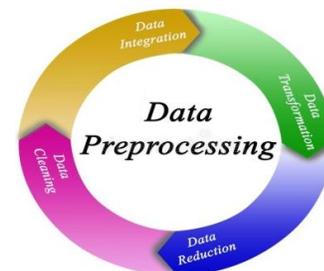


Fig. 3. Data Preprocessing Steps.

Data Reduction for Data Quality: In order to maintain data integrity, we needed to deal with all rows containing null value; hence we opted for python library tool called PyCaret. We have two options which are either to automatically fill all the null values or to remove any row(s) containing null values weighted. Having weighted the risk involves in both, we decided to remove any row with null or empty value from the data, and this was done using PyCaret python library.

Data Transformation: For us to make our data to acceptable format for easy data mining and pattern recognition, it needed to undergo some data transformation. To ensure data is fully transformed to acceptable format, we used Discretization, normalization, and data aggregation technique.

Data Mining: Having successfully set up our apparatus which includes IBM cloud object storage facility, running background window service, active web service, scheduler, and with the data being thoroughly pre-processed and transformed.

Unlike current data mining in cloud computing process in which, one needs to make direct connect to the data warehouse or called the csv file. We only called our web API which was developed and hosted on the cloud as seen below;

Our Artificial intelligence powered anti-cyberbullying system is based on the system of intercepting an outgoing message from the sender. It automatically intercept an incoming message before getting to the receiver and validate the status whether it is bullying or non-bullying. If it is a bullying message, it is automatically blocked from getting to the receiver, and the status of the message will be displayed as 'not delivered' to the sender using a natural language processing mechanism of artificial intelligence. We use machine learning algorithm of Multinomial Naïve Bayes (MNB) and Linear Support Vector Machine (SVM) to train our model, and then expose it to a restful API for global coverage.

A. Implementation with Multinomial Naïve Bayes

Multinomial Naïve Bayes is based on the binomial distribution which is derived from series of combinatorial theorem (Fig. 4).

$$P(X_i|X_j) = P(X_i) \text{ for any distinct } X_i \text{ and } X_j \text{ as well as } P(X_i|X) = P(X_i) \text{ for any } X \subset X \setminus X_i.$$

Expansion and product simplification of the chain rule:

$$P(C_k, X_1, \dots, X_n) = P(C_k) \prod_{i=1}^n P(X_i|C_k)$$

In order to prevent the web service from being overwhelm due to multiple calling of the endpoint at regular interval, we developed a background window service to support the restful API service. Both the web API and the window service are picking records from the integrated data warehouse and pushing to the CSV file on the IBM Object Cloud Storage facility. They automatically pick new records to the CSV, and if a record is modified, it will be picked and modified on the CSV as well. The essence of the scheduler which was written in python is to be calling the web API at regular interval to check and push from the integrated data warehouse to the CSV file.

With the successful setting up of cloud environment and the necessary software programs being in fully execution, we

proceeded by adopting the following data cleansing and preparation techniques.

Data cleaning: We use preprocessing and cleaning methods to remove incomplete data that might cause system failure and also affect output prediction. Rows containing missing values where completely removed. We also use different methods to identify and remove noisy data, outliers, and other factors which can influence the output result.

Data Reduction for Data Quality: In order to maintain data integrity, we needed to deal with all rows containing null value; hence we opted for python library tool called PyCaret. We have two options which is either to automatically fill all the null values or to remove any row(s) containing null values weighted. Having weighted the risk involves in both, we decided to remove any row with null or empty value from the data, and this was done using PyCaret python library.

Data Transformation: For us to make our data to acceptable format for easy data mining and pattern recognition, it needed to undergo some data transformation. To ensure data is fully transformed to acceptable format, we used Discretization, normalization, and data aggregation technique.

B. Implementation with Linear Support Vector Machine

Similar to SVC with parameter kernel='linear', but implemented in terms of liblinear rather than libsvm, so it has more flexibility in the choice of penalties and loss functions and scale better over a large numbers of samples. This class supports both dense and sparse input and the multiclass support is handled according to a one-vs-the-rest scheme (Supervised learning, [online] [11]). We use the inbuilt C implementation which is based on random number generator to select best features so as to ensure low bias and high variance (Fig. 5).

C. Validation and Cross Validation

We obtain an accuracy of over 92% for both Multinomial Naïve Bayes and Linear support Vector Machine, But having such accuracy can be reason for over fitting a times. So we needed to see how the model will behave when fed or exposed to an unfamiliar data. Hence, we didn't jump to conclusion immediately; we decided to do validation and cross validation. We started our validation with confusion matrix and F1 score (Fig. 6).

```
import seaborn as sns
import pandas as pd
import numpy as np
import nltk
from nltk.tokenize import word_tokenize
from nltk import pos_tag
from nltk.corpus import stopwords
from nltk.stem import WordNetLemmatizer
from sklearn.preprocessing import LabelEncoder
from collections import defaultdict
from nltk.corpus import wordnet as wn
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn import model_selection, naive_bayes, svm
from sklearn.metrics import accuracy_score

url='C:\data\data-1\toxicity_parsed_dataset.csv'
twenty_train=pd.read_csv(url)

twenty_train.columns=[twenty_train['text'],twenty_train['oh_label']]

x=twenty_train['text']
y=twenty_train['oh_label']

from sklearn.feature_extraction.text import CountVectorizer
count_vect = CountVectorizer()
X_train_counts = count_vect.fit_transform(x)
X_train_counts.shape

count_vect.vocabulary_.get('algorithm')
```

Fig. 4. Python Code Snippet with Multinomial Naïve Bayes.

```
In [76]: import numpy as np
url='C:\\data\\data-1\\toxicity_parsed_dataset1.csv'
twenty_test = pd.read_csv(url)
docs_test = twenty_test.text
predicted = text_clf.predict(docs_test)
np.mean(predicted == twenty_test.oh_label)

Out[76]: 0.924489201178642

In [77]: from sklearn.linear_model import SGDClassifier
text_clf = Pipeline([('vect', CountVectorizer()), ('tfidf', TfidfTransformer()), ('clf', SGDClassifier(loss='hinge', penalty='l2',
text_clf.fit(twenty_train.text, twenty_train.oh_label)

predicted = text_clf.predict(docs_test)
np.mean(predicted == twenty_test.oh_label)

Out[77]: 0.923938736521711

In [78]: from sklearn.linear_model import SGDClassifier
text_clf = Pipeline([('vect', CountVectorizer()), ('tfidf', TfidfTransformer()), ('clf', SGDClassifier(loss='hinge', penalty='l2',
text_clf.fit(twenty_train.text, twenty_train.oh_label)

Out[78]: Pipeline(steps=[('vect', CountVectorizer()), ('tfidf', TfidfTransformer()), ('clf', SGDClassifier(alpha=0.001, max_iter=5, random_state=42, to=10000))])])
```

Fig. 5. Python Code Snippet with Support Vector Machine.

```
In [82]: from sklearn.metrics import classification_report, confusion_matrix, accuracy_score
print("-----")
print(confusion_matrix(twenty_test.oh_label, predicted))
print("-----")
print(classification_report(twenty_test.oh_label, predicted))
print("-----")
print(accuracy_score(twenty_test.oh_label, predicted))
print("-----")

[[28189  0]
 [ 2149 346]]

-----
precision  recall  f1-score  support
0         0.92    1.00    0.96    28189
1         1.00    0.13    0.23     2149

accuracy
macro avg  0.96    0.56    0.59    30883
weighted avg 0.93    0.92    0.90    30883

-----
0.923938736521711
```

Fig. 6. Validation Report.

With an accuracy of 96% and a very low mean square error (MSE), our model is certain to have low bias and high variance, hence, we proceeded to developing a restful API using Python’s flask framework and then expose our model to the API to ensure the widest coverage. We developed a chatbot automation system and exposed it to our newly developed restful API service. The chatbot automation system has a graphical user interface (GUI) for human interaction.

Since, we have exposed our restful API to our model, the API is automatically called and it immediately intercepts both outgoing and incoming messages and categorizes them based on bullying and non-bullying. If the outgoing message falls into bullying, it automatically filtered it and ensures the receiver doesn’t receive it. The sender too will see from the graphical interface that the message did not deliver. In this case, the bullying message does not reach the receiver (Fig. 7).

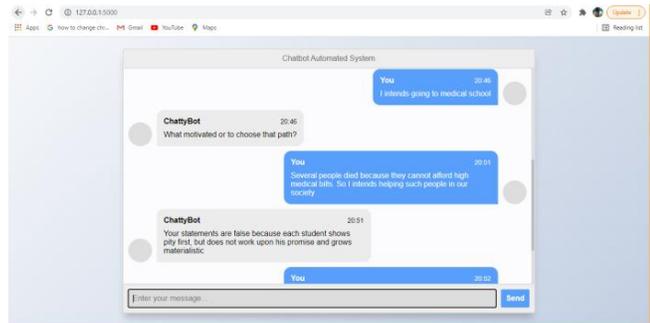


Fig. 7. Chat Messaging GUI.

D. Additional Observation and Time Delivery Comparison

Since, there is an Artificial Intelligence Powered backend implementation that intercepts all outgoing and incoming messages, we deem it necessary to estimate the time for the message to reach the receiver and vice versa since there is an additional backend process of interception. To do this we first removed the backend interception from the newly developed chatbot automation system, and then begin to interact by sending messages from the interface. We discovered that the messages were instantly delivered. Then we decided to re-introduce our restful API which will intercept all messages and feed our trained model, we monitor if for some time during message interaction from the surface. We find out the messages was also delivered on time if it is non-bullying, but if it is bullying the message is not delivered at all. Since, there is no apparent different in time taken when we removed the backend interception process and when we added. We came to conclusion that the extra time taken by the backend message interception process which determine whether the message should be sent to the receiver or not is fraction of a seconds which is negligible.

IV. CONCLUSION

With our research and implementation, we believed the problem of cyber bullying will be solved once and for all if our implementations can be integrated in all the online social media and messengers. Validation and cross validation of our results shows accuracy of 92%, with low bias and high variance and also a very low mean square error (MSE). Our implementation can be adopted and used on any real life social media or online messenger.

Our conclusion will be incomplete without mentioning the limitation of our developed model. The only limitation to our model is the data. The more the data used to train model, the more accuracy it will be.

The data used for this research is from different social media platforms like Kaggle, Twitter, Wikipedia Talk pages and YouTube. The data contain text and labeled as bullying or not. The data contains different types of cyber-bullying like racism, hate speech, aggression, insults and toxicity.

The greatest breakthrough in machine learning is going to be when a deployed model can automatically retrain itself by constantly getting live data from a source, train and update itself without any human effort. Since, model accuracy is somehow dependent with availability of data, and millions of new data are constantly being available in the cloud on daily bases after deploying our model. This will happen at some later time and we will be able to have a super model that is able to get live data, retrain, and redeployed itself on daily bases without human intervention.

REFERENCES

- [1] S. A. Hemphill, A. Kotevski and J. A. Heerde, "Longitudinal associations between cyber-bullying perpetration and victimization and problem behavior and mental health problems in young australians", International journal of public health, vol. 60, no. 2, pp. 227-237, 2015.
- [2] R. S. Tokunaga, "Following you home from school: A critical review and synthesis of research on cyberbullying victimization", Computers in human behavior, vol. 26, no. 3, pp. 277-287, 2010.

- [3] S. Qaiser and R. Ali, "Text mining: use of tf-idf to examine the relevance of words to documents", *International Journal of Computer Applications*, vol. 181, no. 1, pp. 25-29, 2018.
- [4] S. Tong and D. Koller, "Support vector machine active learning with applications to text classification", *Journal of machine learning research*, vol. 2, pp. 45-66, Nov 2001.
- [5] A. Perera and P. Fernando, "Accurate cyberbullying detection and prevention on social media", *Procedia Computer Science*, vol. 181, pp. 605-611, 2021.
- [6] R. R. Dalvi, S. B. Chavan and A. Halbe, "Detecting a twitter cyberbullying using machine learning", 2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS), pp. 297-301, 2020.
- [7] S. Paul, S. Saha and M. Hasanuzzaman, "Identification of cyberbullying: A deep learning based multimodal approach", *Multimedia Tools and Applications*, pp. 1-20, 2020.
- [8] C. Iwendi, O. Srivastava, S. Khan and P. K. R. Maddikunta, "Cyberbullying detection solutions based on deep learning architectures", *Multimedia Systems*, pp. 1-14, 2020.
- [9] Y. Liu, P. Zavorsky and Y. Malik, "Non-linguistic features for cyberbullying detection on a social media platform using machine learning", *International Symposium on Cyberspace Safety and Security*, pp. 391-406, 2019.
- [10] Stemmers, [online] Available: <https://www.nltk.org/howto/stem.html>.
- [11] S. Rahman, K. H. Talukder and S. K. Mithila, "An Empirical Study to Detect Cyberbullying with TF-IDF and Machine Learning Algorithms," 2021 International Conference on Electronics, Communications and Information Technology (ICECIT), 2021, pp. 1-4, doi: 10.1109/ICECIT54077.2021.9641251.
- [12] K. Rizwan, S. Babar, S. Nayab and M. K. Hanif, "HarX: Real-time harassment detection tool using machine learning," 2021 International Conference of Modern Trends in Information and Communication Technology Industry (MTICTI), 2021, pp. 1-6, doi: 10.1109/MTICTI53925.2021.9664755.
- [13] Dinakar, K., Reichart, R., & Lieberman, H. (2021). Modeling the Detection of Textual Cyberbullying. *Proceedings of the International AAAI Conference on Web and Social Media*, 5(3), 11-17. Retrieved from <https://ojs.aaai.org/index.php/ICWSM/article/view/14209>.
- [14] Tomkins, Sabina & Getoor, Lise & Chen, Yunfei & Zhang, Yi. (2018). A Socio-linguistic Model for Cyberbullying Detection. 53-60. 10.1109/ASONAM.2018.8508294.
- [15] Pawar, R., Agrawal, Y., Joshi, A., Gorrepati, R., & Raje, R. R. (2018). Cyberbullying Detection System with Multiple Server Configurations. 2018 IEEE International Conference on Electro/Information Technology (EIT), 0090-0095. <https://doi.org/10.1109/EIT.2018.8500110>.

Replica Scheduling Strategy for Streaming Data Mining

Shufan Li¹, Siyuan Yu², Fang Xiao³

Computer Science and Artificial Intelligence School, Wuhan University of Technology, Wuhan, China^{1,2}
Wuhan National Laboratory for Optoelectronics, Huazhong University of Science and Technology, Wuhan, China³

Abstract—In a distributed storage and computing framework, traditional streaming data mining techniques are inefficient when processing massive amounts of data. In this paper, we take the copy in cloud storage as an allocatable resource for scheduling and propose a RepRM strategy to improve the efficiency of data mining and analysis. The key idea of this work is to take the data copy as the resource to be allocated, and use the backward inference method of dynamic programming to solve the data copy ratio, the optimal number of copies is obtained. Experiments and observations have proved that compared with the traditional scheduling method of Hadoop, after adopting the RepRM strategy scheduling, the memory resources of the homogeneous cluster are saved by about 40-50% during parallel mining of streaming data, and the throughput rate is increased by 20% to 30%.

Keywords—Streaming data mining; dynamic programming; replica scheduling strategy; cloud computing

I. INTRODUCTION

The continuous development of computer science has resulted in more and more tasks and data needed to be processed, and the computer processing capacity and processing speed have been difficult to meet the needs of users. As a product of virtual technology, cloud computing can process massive data and tasks. However, due to the huge amount of computing, the cloud platform needs to allocate the system resources to each task in the computing process reasonably. What's more, more and more streaming data are stored in the cloud, especially large data analysis systems store a great deal of data, such as data generated by sensors, generated by network management equipment, and generated by core switches, such as log data, audio and video data, etc. and these data are continuously generated in a data stream according to time. Consequently, an effective strategy to allocate the system resources to each task in the computing process is vital.

The mining of massive streaming data is applied to various fields and it produces valuable analytical results for it. Taking marketing as an example, by mining and analyzing the market behavior of a large number of user data, we can guide the further market working by getting the consumption habits of users. For example, according to the consumption situation of users' credit cards, we can directly know the main consumer needs, shopping interests and consumption concepts, which is very valuable information for marketing and product promotion, and is helpful to guide the next market planning.

Because massive streaming data has the characteristics of big data, and for the upper application of cloud storage, the processing and analysis of massive streaming data are very different from the previous processing. Taking data query and mining as examples, users need to query and analyze the accumulated data for a long time when analyzing the data, which has high requirements for the data searching and comparison performance in cloud storage, while the past data query only queries and compare for a certain data. So the traditional streaming data mining technology is inefficient when processing huge amounts of data. Uncontrollable and continuous surge of large volumes of data has exacerbated the trend of "the explosion of data and the lack of knowledge". At this time, the distributed computing platform has become a research hotspot as an effective means to solve the problems [1,2].

Take Hadoop as an example, the homogeneous or heterogeneous distributed computing platform built by it can meet the needs of large-scale data processing technology in scientific research, engineering and other fields, it has become one of the mainstream data processing frameworks of large Internet companies at home and abroad, such as the data processing applications of Yahoo, Twitter, and other companies. With the increasing scale of the Hadoop cluster and the increasing fields of use, the management and usage of resources are increasingly valued. Many researchers take YARN's resource scheduling algorithm as their research direction, through researching and designing reasonable resource allocation algorithms to achieve higher resource utilization. But among the numerous studies, according to the characteristics of streaming data, it is rare to design and implement the resource scheduling strategy of the data mining algorithm, so that the current mining model for streaming data in the cloud platform cannot effectively allocate the resources in the cluster, or complete the mining of streaming data efficiently.

In practical applications, most streaming data mining algorithms are aimed at a certain type of streaming data [3,5]. The operating parameters of the algorithm are determined in the program initialization stage and cannot be dynamically modified, which is called a "static algorithm" [4]. Although some algorithms can be regulated through parameter adjustment to adapt to the dynamic application environment, these dynamic algorithms are relatively rigid and have no learning and adaptive capabilities. For example, Franke pointed out that streaming data is usually a sequence of data objects with time as the latitude. Therefore, the execution process of

more streaming data mining algorithms is sorted according to the time when the data arrives sequentially, and a single linear scan method is used to analyze the data. Analyze and compare [4].

Alternatively, the relationship between steaming data and environmental differences has been considered by Knorr and other professors. Different environments can lead to differences in the expression form of streaming data, which can then lead to differences in data mining algorithms. For example, the resource allocation of a computer system is dynamic, and the data mining algorithms may allocate more or fewer resources (such as CUP, memory, etc.) because of the actual load situation of the computer system, and the number of resources will directly affect the response time and processing time of the streaming data mining algorithm. Therefore, an adaptive resource scheduling method is needed so that the resources can be adjusted according to the different streaming data.

The work done in this paper is as follows: First, verify the impact of replicas on the throughput of streaming data mining. After analyzing the relationship between the copy and the throughput rate of streaming data mining, the resource allocation model of streaming data mining is established. Secondly, according to the characteristics of streaming data, the resource scheduling model of the cloud platform incorporates the copy data resources and the resource (replica) scheduling model for the cluster. Once again, build a model to analyze and solve. Finally, based on the Hadoop big data platform, the scheduling model was implemented by improving the YARN component. What is more, in order to verify the effectiveness of the copy-based resource scheduling model, we use two types of test data sets and real network traffic data sets to test the improved resource scheduling strategy.

II. RELATED WORK

In order to propose an efficient resource scheduling model, we should consider the challenges which come from several aspects. Then we will investigate the recent research about these challenges as follow to build a proper model [6].

Dominant Resource Fairness (DRF) is a multi-resource allocation algorithm of max-min fairness [7]. The authors introduced that dominant share is the largest share of any resources that distribute to the user, and DRF conceived that the number of resources assigned to the user should be determined by its dominant share. After all, the purpose of DRF is to seek the maximum of the minimum share among all users. While DRF figured out the demand heterogeneity of multiple resources, however, DRF neglects the heterogeneity, it based on an assumption that computing resources are generally assumed to be isomorphic. In order to resolve more problems, many researchers have optimized and expanded DRF. DRFH [8] is a multi-resource allocation mechanism. For DRFH, the resources are pooled by a large number of heterogeneous servers; they are representing processing, memory, storage and so on. DRFH equalizes the global dominant share allocated to each user. However, it fails to consider users' placement constraints. PS-DSF [9], an extension of DRF is applicable for heterogeneous resource-

pools in the presence of placement constraints. Its solution is defining a virtual dominant share.

For further study, the current schedulers such as DRF which only for the fairness can't meet our needs, so professionals investigated other schedulers which consider the counterbalance between performance and fairness [6]. Gemini [10] considers two scheduling policies during runtime, the former maximizes the utilization of resources by balancing the remaining capacity of the resources of the node, while the latter is to fairly allocate resources to users in a system containing multiple types of resources. During the adjustment process, the strategy will be selected by estimating the loss of performance and fairness.

The cost of a great deal of energy consumption has taken up a considerable part of the total cost of the data center. When we not only need to reduce this part of the cost, but also meet the QoS requirements of users, it's been a main research topic of resource scheduling strategy. EFS [11] is an Energy-aware Fair Scheduling framework based on YARN, EFS uses dynamic node management. It can meet the users' QoS requirements and reduce the energy consumption, because of the energy-aware resources scheduling and the strategy of turning off the unused nodes for a specified duration. However, EFS does not consider the data distribution and replication dependencies.

In recent years, cloud storage systems have emerged as a promising technology for storing data blocks on various cloud servers [12], and replica is the basic means of tolerance and availability of the distributed systems, so the distribution of data copy is significant to the systems [13], many problems can be resolved by data replication algorithms. However, data replication also produces energy consumption and costs, so it is very important to reasonably schedule copy resources in resource scheduling, there are also many kinds of research about this topic.

In order to solve the problem about the placement of data replica, Cui et al. [14] build a tripartite graph-based model, and propose a genetic algorithm-based data replica placement strategy. The dynamic multi-objective optimized replica placement and migration strategies for SaaS applications in edge cloud are proposed by Chunlin et al. [15]. According to the result of the experiment, this strategy can improve the utilization rate of network resources. Huang and Wu [16] not only proposed an optimization model for data replication and placement problem, but also designed hybrid genetic algorithm based on data support degree to solve the model. The algorithm is found to have good performance by using real data set. Khojant et al. [17] proposed Predictive Frzyzy Replication (PFR). The new algorithm can replicate the historical usage of files, files size, the level of the sites and free available space for replication in advance and decide which replications should be deleted through forecasting of future demand and the relevant file of the replications to save cost. Salem et al. [18] created a new algorithm derived from a combination of ABC and Multi-Objective Optimization. The proposed MPABC algorithm enables fast access to the data and selects the best copy location closest to the user.

In these present studies, the purpose of most resource scheduling models is to allocate resources such as CPU, memory and bandwidth, they don't consider the data replica, CPU and memory simultaneously, unified scheduling. So we incorporate data copy as a scheduling resource when build a new resource scheduling model. And from the perspective of the whole process of data processing, the bottleneck of the big data processing performance in the distributed computing platform lies in the data transmission consumption, not the CPU computing power [19]. Therefore, an effective resource scheduling method is needed to select appropriate nodes for data processing, that is, considering the location of the current copy, data mining is performed on the node where the copy is located, so that the mining speed is greatly improved. Based on this, this article starts from the perspective of copy selection, takes the copy as a data resource and considers it as the computing resource at the same time, and uses the dynamic programming model to design the resource scheduling model of the big data platform to improve the data mining throughput rate of the cloud platform.

III. MATERIALS AND METHODS

A. Distributed Application and Copy Selection

In order to take Hadoop and OpenStack as examples, the copy in the cloud platform is generally set to 3, which means that the same data may only exist in a few nodes of the cloud platform. Therefore, when a distributed application applies for the use of resources such as CPU and memory, the positional relationship between the node where the resource is located and the copy is extremely important.

1) *Verification of the impact of replicas on distributed mining:* For the purpose of verifying and testing the impact on data copies on distributed mining, we use the KDD CUP 2000 data set for verification, and use Hadoop clusters and stand-alone machines to analyze and mine streaming data respectively. Table I shows that the Hadoop and single configuration used for the experiment. The experimental process is to extract NetFlow seven-tuples from KDD CUP 2000. What is more, The Hadoop test uses MapReduce to submit a task which is NetFlow analysis to Hadoop, and the stand-alone test uses the network interactive analysis tool set SILK for NetFlow seven-tuple analysis.

The results of the experiment are shown in Fig. 1. It can be found from the experiment that when the total amount of data is small (less than 400G), in contrast to the Hadoop distributed platform, in data analysis ability, a single machine has apparent advantages, which is just 70%. When the total amount of data is large (greater than 400G), the Hadoop distributed platform shows its advantages progressively. When the amount of data is 500G, the analysis time is only 84% of the single machine. But 5 servers make up this distributed platform, and the single platform uses just one server. It can be seen that using the single platform used in the experiment can process 400GB stream data within 15 minutes, which can meet the seven-tuple analysis of about 3.5Gpbs network traffic.

TABLE I. SERVER CONFIGURATION OF THE EXPERIMENT

Hadoop	Stand-alone
4 slave hosts: CPU: Two-way four-core 2.6G RAM: 48G Disk: 40TB	CPU: Two-way eight-core 2.6G RAM: 60G Disk: 20TB
1 master host: CPU: Two-way eight-core 2.6G RAM: 60G Disk: 20TB	

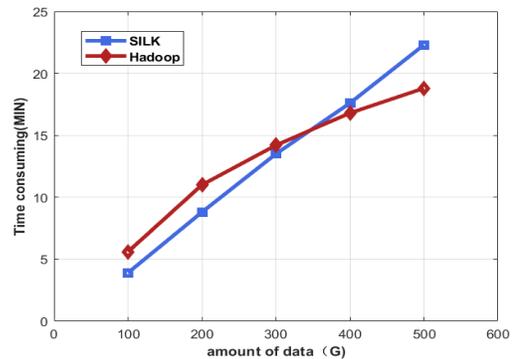


Fig. 1. Time Consuming (MIN).

In situation of a bit of data, the location of the replica in the Hadoop cloud platform will affect the throughput rate of the distributed platform for traffic analysis. Hadoop's three-copy strategy requires some computing nodes to wait for the completion of the network transmission of data, thus increasing the time consumption of network transmission; and the MapReduce task allocation mechanism has a certain time loss at the beginning and end of the task. Therefore, in the case of a relatively small amount of data, low throughput for traffic analysis on distributed platforms.

2) *The optimization problem of data copy in data mining:* Through the verification of 1), disposable resources such as replicas are important factors that apparently affect the capability of data mining algorithms. Both the results and throughput of data mining algorithms not only depend on the allocation of computing hardware resources such as CPU time and available memory, but are also closely related to the number of copies. In the case of a small number of copies and limited computing resources, there will be a disparity between data mining algorithm results and the optimal results.

The mining model can adjust itself by modifying its copy resources and computing resources, thereby increasing the throughput of the mining model. Therefore, to make the throughput of the mining algorithm higher, we need to adjust the computing resources's allocation and data resources. Especially for real-time streaming data mining, it is more necessary to dynamically adjust how computing resources and data resources are allocated, so that the real-time mining model is resource-adaptive [4].

Take the WEB server as an example, the WEB server needs to process a large number of incoming and outgoing data packets in the network in real-time. This is the typical stream of data. Within a period of time, the WEB server will receive data packets from the Internet and send a certain number of data packets to the Internet in chronological order. These data packets are usually organized in the form of IP data packets. Each IP data packet constitutes a stream data object, which together constitutes a data stream on the WEB server. When the network traffic is large enough, due to the limited resources of the host, online analysis, offline analysis, and initial filtering of stream data will all cause conflicts in resource usage.

For the contradiction between resources and speed, the solution is to effectively allocate limited data resources and computing resources [20, 21]. Taking the traditional mining model as an example, it will affect the four modules of the model: data filtering module, online mining module, offline mining module and resource detection module. Assuming that the process of processing a stream data object includes data online mining module, resource detection module, offline processing module and data filtering module, the total amount of available resources is R , of which data online mining module, resource detection module, offline processing module and the resources consumed by the data filtering module are R_1 , R_2 , R_3 and R_4 . Among these four modules, the online mining module and the data filtering module need to detect and process the data in real-time, so the resources consumed are increased or decreased at the same time; on the other hand, the offline mining module processes the data offline, the resources which are consumed do not increase or decrease at the same time as the first two. According to this situation, the resource consumption can be adjusted into three parts: the overall consumption R_1+R_4 of the online mining module and the data filtering module, and the consumption R_2 and R_3 of the resource detection module and the offline mining module. So when the total amount of resources R is certain, how to effectively allocate the resources of R_1+R_4 , R_2 and R_3 , so that the throughput rate of the online mining module, resource detection module, offline mining module and data filtering module can be increased to V_1 , V_2 , V_3 and V_4 , which maximizes the throughput rate, is a dynamic programming problem that optimizes resource allocation. To resolve this problem, the following modeling is required:

$$R = \sum_{i=1}^n R(p, z(p)) \quad (1)$$

In order to deal with the problem about dynamic programming, the resource allocation process is first divided into n different stages to allocate resources. In the p th stage, the p th module is allocated resources by the system. After the allocation is completed, the number of resources remaining in the system is $q(p+1)$. Then the system allocates the remaining resources $q(p+1)$ to the $p+1$, $p+2$, ..., n th modules. The optimal function value that can be obtained in the $p+1$ stage is set to $t(q(p+1), p+1)$, that is, under the premise of resource $q(p+1)$,

the final $p+1$ to n is completed. The maximum mining throughput rate that can be obtained by the allocation of each module.

Then, for dynamic programming problem, the basic equation is:

$$t(p, q(p)) = \max_{z(p) \in M(q(p))} \{R(p, z(p)) + t(q(p+1), p+1)\},$$

$$p = n-1, \dots, 1 \quad (2)$$

$$t(n, q(n)) = R(p, z(n))$$

The corresponding state transition equation is:

$$q(p+1) = q(p) - z(p) \quad (3)$$

For the resource optimization problem, according to Formula 2 and Formula 3, a related mathematical model is established. To solve the model, the optimal decision sequence $(z'(1), z'(2), \dots, z'(n))$ can be obtained through the inverse method, and then the maximum throughput rate $R(q(1))$ of the mining algorithm can be obtained. Specifically, in the data mining model, n is the 4 modules in the model, that is, $n=4$, and the total resource number S is the number of copies of the data resource.

In a distributed system, it is supposed that the resources of S units need to be allocated to n modules, where S represents the number of stream data objects that can be processed and is a positive integer. Assuming that the throughput of data mining can be increased to $R(k, z(k))$ after $z(k)$ resource units are allocated to the k th module, then the overall goal of resource allocation is to allocate S to each module. The resources of each unit finally make the total throughput of mining reach the highest, that is, the R value in Formula 1 reaches the maximum.

In this paper, we know that in the distributed framework, data mining is usually carried out in parallel, so it is usually multiple mining algorithms to mine massive data. Taking a typical parallel mining model as an example, suppose that there are S units of resources need to be allocated by distributed system to n parallel mining algorithms, and S is the number of data copies that can be allocated. Assuming that the throughput rate of data mining can be increased to $R(p, z(p))$ after allocating $z(p)$ units of resources to the p -th mining algorithm, then the overall goal of resource allocation is to be reasonable for the data mining algorithm to allocate S data copies, and finally maximize the mining throughput rate, in other words, the R value in Formula 1 reaches the maximum.

Similarly, in order to deal with the optimization problem, the allocation replica is grouped into n distinct phases in the resource allocation process in a distributed system. The p th algorithm is allocated data copy resources by the system in the p th stage. After the allocation is completed, the number of copies remaining in the system is $q(p+1)$. Similarly, the optimal function value that can be obtained in the $q(p+1)$ stage is set to $t(q(p+1), p+1)$. According to Formula 2 and the related theory of dynamic programming, the basic equation of the mathematical model can be determined at first:

$$t(p, q(p)) = \max_{z(p) \in M(q(p))} \{R(p, z(p)) + t(q(p+1), p+1)\},$$

$$p = n-1, \dots, 1$$

$$t(n, q(n)) = R(p, z(n)) \quad (4)$$

According to the above Formula 4, the state transition equation in the mathematical model can be determined as:

$$q(p+1) = q(p) - z(p) \quad (5)$$

According to the established mathematical model, the solution of this model can be used to obtain the optimal decision sequence $(z'(1), z'(2), \dots, z'(n))$ through the inverse method, so as to the mining throughput reaches the maximum value $R(q(1))$.

B. Replica Scheduling Strategy RepRM for Streaming Data

For big data systems (such as Hadoop systems), the distributed resource scheduling problem is an NP-hard problem. Mass flow data has a certain degree of no aftereffect, so the traditional scheduling algorithm can't allocate system resources well. In the whole process of mining algorithm operation, if adjustment of resources is a decision, the data mining between two decisions is a stage. Therefore, the resource allocation problem in distributed mining of streaming data is a dynamic programming problem. When performing distributed mining on streaming data, the resource allocation for it can be described as: how to allocate limited resources to multiple mining algorithms, so that the mining model can mine the streaming data to the greatest extent.

For various algorithms of parallel mining, take K-Means, KNN and Apriori algorithms as examples. It is assumed that each algorithm is performing streaming data mining, we need to consider how to build a model which can allocate limited resources to these algorithms to make the entire parallel mining model throughput rate be the largest. The problem comes down to how to allocate data copy resources to maximize the throughput of the parallel mining model, which is also a resource scheduling problem.

1) *Replica scheduling based on the copy*: Assuming that there are R assignable data copy resources in the distributed computing platform, n data mining algorithms are running on the platform at the same time, and the throughput of the algorithm in mining is related to the amount of replica resources put into use. Assuming that the i-th data mining algorithm are allocated by r_i data copy resources, and the throughput rate of the i-th mining algorithm is $f_i(r_i), i = 1, 2, 3, \dots, n$. At this time, the whole throughput rate of the platform's n algorithms on mining is $f_i(r_i), i = 1, 2, 3, \dots, n$.

Then the problem boils down to how to allocate R data copy resources: for n data mining algorithms, in order to maximize the total throughput rate, the total throughput rate reaches $g_n(r_n)$. The programming model is as below:

$$\text{Max} : Z = f_1(r_1) + f_2(r_2) + \dots + f_n(r_n)$$

$$\text{s.t.} \begin{cases} r_1 + r_2 + \dots + r_n = R \\ r_i \geq 0, i = 1, 2, \dots, n \end{cases} \quad (6)$$

This depends entirely on the throughput function $f_i(r_i)$ of each mining algorithm. When $f_i(r_i)$ is a linear function. It is a linear programming problem; when $f_i(r_i)$ is a non-linear function, it is a nonlinear programming problem. Especially when n is relatively large, the solution process is extremely troublesome. However, due to the inefficiency of massive data, the problem is to solve a parallel resource allocation model, which can be solved by using the inverse relationship of dynamic programming.

Let the state variable of the number of data replication resources allocated to algorithm k to algorithm n be s_k , and the decision variable u_k represents the number of data copy resources allocated to algorithm k. After the data copy resource allocation decision of algorithm k is completed, let the number of data copy resources obtained by algorithm k be r_k , that is, $u_k = r_k$ after the allocation is completed. At this time, the number of replica resources r_k allocated to algorithm k satisfies:

$$s_{k+1} = s_k - r_k \quad (7)$$

And because $u_k = r_k$, you can get:

$$s_{k+1} = s_k - r_k = s_k - u_k \quad (8)$$

The allowed decision set is:

$$D_k(s_k) = \{u_k \mid 0 \leq u_k = r_k \leq s_k\} \quad (9)$$

When the number of data copy resources of s_k is allocated to the k-th to n-th algorithms, the platform can get the maximum mining throughput rate of $g_k(s_k)$, and the inverse relationship of dynamic programming can be obtained:

$$\begin{cases} g_k(s_k) = \max_{0 \leq r_k \leq s_k} \{f_k(r_k) + g_{k+1}(s_k - r_k)\}, k = n-1, \dots, 1 \\ g_n(s_n) = \max_{r_n = s_n} f_n(r_n) \end{cases} \quad (10)$$

Using Formula 10, we can calculate one by one algorithm, and finally get $g_1(s_1)$.

2) *Replica scheduling model RepRM*: Distributable data copy resources and changes in data characteristics are two important factors that affect the throughput of data mining algorithms. In distributed streaming data mining, the available resources can be reasonably allocated to the mining model according to the number of copies and the characteristics of the streaming data, and the throughput rate of the model can be improved.

If only considered the computing resources, the CPU and memory contained in each server are almost the same, and the allocatable resources used by each node for streaming data mining are also almost the same. Then the problem can be simplified to the configuration of data copy resources. At this point, the problem turned into a resource allocation problem.

Take a Hadoop cluster as an example. The cluster is composed of R identical servers (for example, DELL 820), which are used for network traffic analysis and mining. When multiple different users submit n network traffic data mining algorithm requests, the mining throughput rate is $f_i(r_i), i=1,2,3,\dots,n$, and the inverse method can be used to solve the problem. Assume that in this example, the cluster is composed of 5 homogeneous servers. At present, the Apriori algorithm is already running for online mining of streaming data. At the same time, two users request to use the KNN algorithm and the K-Means algorithm for offline mining of streaming data. And according to the number of data copies allocated to the mining algorithm, the sampling size during mining is also inconsistent. The specific sampling size is shown in Table II.

TABLE II. DATA COPY RESOURCE ALLOCATION AND SAMPLING SIZE OF HOMOGENEOUS CLUSTER

Number of servers	0	1	2	3	4	5
K-Means	0 Gbps	3 Gbps	7 Gbps	9 Gbps	12 Gbps	13 Gbps
KNN	0 Gbps	5 Gbps	10 Gbps	11 Gbps	11 Gbps	11 Gbps
Apriori	0 Gbps	4 Gbps	6 Gbps	11 Gbps	12 Gbps	12 Gbps

At this time, according to Formula 10, the inverse method can be used to obtain the optimal solution, and the solution process is as follows.

Solution:

$$\begin{cases} g_k(s_k) = \max_{0 \leq r_k \leq s_k} \{f_k(r_k) + g_{k+1}(s_k - r_k)\}, k = n-1, \dots, 1 \\ g_n(s_n) = \max_{r_n = s_n} f_n(r_n) \end{cases} \quad (11)$$

Stage 3: When assigning s_3 data copies ($s_3=0,1,2,3,4,5$) to the Apriori algorithm, the sampling size is:

$$g_3(s_3) = \max_{r_3} [f_3(r_3)], r_3 = s_3 \quad (12)$$

Because only the Apriori algorithm is mining at this time, all data copies in the cluster can be allocated to it, so its sampling size is the maximum sampling size at this stage, as shown in Table III, in the table r_3^* means that $g_3(s_3)$ is the optimal decision at the maximum value.

Stage 2: Assuming that s_2 data copies ($s_2=0,1,2,3,4,5$) are allocated to the Apriori algorithm and the KNN algorithm, then for each s_2 value, the available sample size is:

$$g_2(s_2) = \max_{r_2} [f_2(r_2) + g_3(s_2 - r_2)], r_2 = 0,1,2,3,4,5 \quad (13)$$

TABLE III. RESOURCE DYNAMIC DECISION OF HOMOGENEOUS CLUSTER STAGE 3

r_3	$f_3(r_3)$						$g_3(s_3)$	r_3^*
	0	1	2	3	4	5		
0	0						0	0
1		4					4	1
2			6				6	2
3				11			11	3
4					12		12	4
5						12	12	5

Because r_2 data copies of the KNN algorithm are allocated, the mining throughput rate is $f_2(r_2)$, and the remaining $s_2 - r_2$ data copies are used in the Apriori algorithm, so the Apriori throughput rate is $g_3(s_2 - r_2)$. The sample size is $f_2(r_2) + g_3(s_2 - r_2)$ at this time. Therefore, it is necessary to select an appropriate value of r_2 to maximize the function. At this time, the numerical calculation table is shown in Table IV.

Stage 1: In order to obtain the maximum mining throughput rate, 5 data copies need to be allocated to the algorithms for calculation. Therefore, when $s_1=5$ data copies are allocated to the Apriori, KNN and K-Means algorithms, the mining throughput rate is:

$$g_1(5) = \max_{r_1} [f_1(r_1) + g_2(5 - r_1)], r_1 = 0,1,2,3,4,5 \quad (14)$$

At this time, the maximum value of this function is the maximum mining throughput rate. The specific numerical calculation table is shown in Table V.

According to the numerical calculation table, there are two executable solutions:

1) When $r_1^*=0$, look up Table V and Table IV to know that the allocation plan at this time is: $r_1=0, r_2=2$, and $r_3=3$.

TABLE IV. RESOURCE DYNAMIC DECISION OF HOMOGENEOUS CLUSTER STAGE 2

r_2	$f_2(r_2) + g_3(s_2 - r_2)$						$G_2(s_2)$	r_2
	0	1	2	3	4	5		
0	0						0	0
1	0+4	5+0					5	1
2	0+6	5+4	10+0				10	2
3	0+11	5+6	10+4	11+0			14	2
4	0+12	5+11	10+6	11+4	11+0		16	1,2
5	0+12	5+12	10+11	11+6	11+4	11+0	21	2

TABLE V. RESOURCE DYNAMIC DECISION OF HOMOGENEOUS CLUSTER STAGE 1

r_1	$f_1(r_1) + g_2(5 - r_1)$						$G_1(5)$	r_1
	0	1	2	3	4	5		
5	0+21	3+16	7+14	9+10	12+5	13+0	21	0,2

That is, the K-Means algorithm does not allocate data copies, the KNN algorithm allocates 2 data copies, and the Apriori algorithm allocates 3 data copies. At this time, the maximum throughput rate of the mining algorithm is 21Gbps.

2) When $r_1^* = 2$, look up Table V and Table IV to know that the allocation plan at this time is: $r_1=2$, $r_2=2$, and $r_3=1$.

That is, the K-Means algorithm allocates 2 data copies, the KNN algorithm allocates 2 data copies, and the Apriori algorithm allocates 1 data copy. At this time, the maximum throughput rate of the mining algorithm is 21Gbps.

IV. RESULT

A. Experiments and Observations

This article implements the RepRM model by revising the resource management and scheduling component (YARN) in Hadoop and modifying the Scheduler component in the Resource Manager module. By default, YARN only supports memory scheduling (such as Capacity strategy, Fair strategy). It uses "containers" to encapsulate memory and CPU. When tasks have resource requirements, they apply to YARN for the CPU and memory "containers" required by the task.

In the testing session, we conducted experiments and observations on the RepRM replica scheduling method. For the parallel streaming data mining model, we used KNN, K-Means and Apriori algorithms to simultaneously mine online. Then observe the differences between YARN's built-in Capacity strategy, Fair strategy and the improved RepRM strategy.

1) *Experimental environment*: In the experiment and observation, the Hadoop cluster used for testing is a 2U rack-mounted DELL PowerEdge FX2 server with a convergent architecture, which contains 4 nodes, and the operating system uses CentOS. The hardware configuration of each node is shown in Table VI. The experimental data includes two test data sets: the test data set generated by the IBM synthesizer and the WEB access traffic data.

a) *Experimental data set: IBM synthetic data*: The data set in this experiment is the T10-I5-D1000K data set produced by the IBM synthesizer [10], where T, I, and D mean the average length of the transaction, the average length of the pattern, and the number of transactions.

b) *Experimental data set: WEB access traffic*: The WEB access traffic used for testing was collected at the network exit of the library of Huazhong University of Science and Technology. The collection program is connected to the egress gateway through the bypass, and the program filters and saves the traffic of the WEB service.

c) *Evaluation index*: In the comparative experiment, YARN's built-in Capacity strategy, Fair strategy, and dynamic planning improved RepRM scheduling strategy are used for resource scheduling, and parallel data mining using K-Means, KNN, and Apriori algorithms. We observe the performance of RepRM from multiple angles: For the mining of test data sets, we compare the throughput and memory usage of parallel mining. By comparing the data obtained from these two sets of

experiments, we observe the different performance of the Capacity strategy, Fair strategy and RepRM strategy.

2) *Mining experiment of IBM synthetic data set*: In experiments based on data sets, we will pay attention to analyze the time complexity and space complexity of the algorithm. When comparing the two data, because the DARPA 99 data collective is at the GB level, and the T10-I5-D1000K is only at the MB level, in order to visualize it on the chart, we will use the test result value of the T10-I5-D1000K data set. It is 1000 times larger, so that it can be compared with the test results of the GB-level DARPA 99 data set.

In the time complexity test, the execution time of the mining model is compared. Fig. 2 shows the parallel mining results of the Capacity strategy, the Fair strategy and the RepRM strategy. It can be seen from Fig. 2 that the execution time of the RepRM strategy is lower than the Capacity strategy, what is more, it is also lower than the Fair strategy that Hadoop comes with. According to the data shown in the experimental results, compared with the Capacity strategy, the execution time of the RepRM strategy is about 70%, and compared with the Fair strategy, the execution time is about 65%. Therefore, the RepRM strategy increases the mining throughput rate by 25% compared with the Fair strategy and 30% compared with the Capacity strategy.

In the comparison of space complexity, the comparison is still based on the memory footprint of the algorithm. The specific data and comparison are shown in Fig. 3. The RepRM strategy has the least memory footprint, the Fair strategy has the middle memory footprint, and the Capacity strategy has the largest memory footprint. The experimental results show us the memory container consumption of the RepRM strategy is only 60%, and only about 53% compared to the Fair strategy.

TABLE VI. SERVER CONFIGURATION OF THE EXPERIMENT

Model	DELL FX2(Including 4 blade servers)
Master	128GBRAM/ 2TBDisk/Operating System CentOS 6.5
Slave1-Slave3	64GBRAM/ 2TBDisk/Operating System CentOS 6.5
Hadoop version	Cloudera 2.2
Network environment	1000M

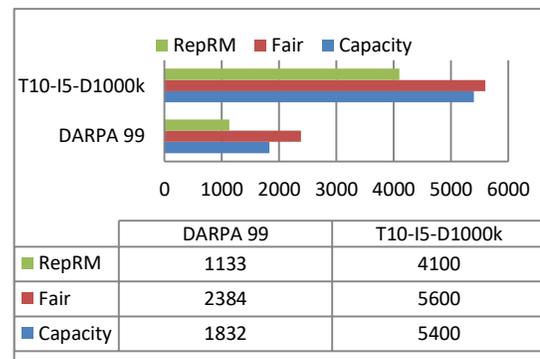


Fig. 2. Execution Time Comparison for the Three Strategies in the Parallel Model.

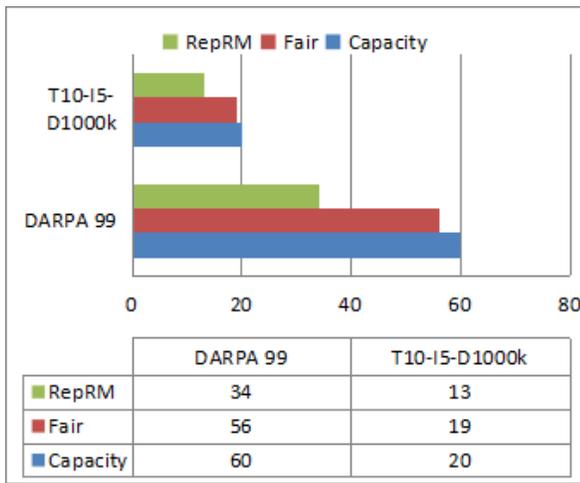


Fig. 3. Memory Footprint Comparison for the Three Strategies in the Parallel Model.

Through four sets of experimental data, it can be seen that the RepRM strategy improves the mining throughput rate by about 25-30% for the resource scheduling strategy that comes with YARN, and saves about 40% of resources in the consumption of memory containers. The reason is:

a) *RepRM strategy is resource-adaptive.* When the Capacity strategy and Fair strategy allocate tasks in Map-Reduce, they only allocate resources based on the existing remaining resources, rather than allocating tasks based on all running tasks. RepRM is based on the throughput of all tasks for dynamic scheduling and appropriate allocation of resources.

b) *YARN's own strategy does not consider the impact of copies on mining throughput.* For distributed mining under the condition of multiple copies, after the task is allocated by Map-Reduce, when there are more computing resources such as CPU and memory, it is necessary to wait for the network transmission of the copy. RepRM allocates computing resources based on the number of data copies, saving network transmission waiting time.

3) *Mining test of network traffic:* Before the start of the experiment, first, we use a data capture tool to capture and save the throughput network traffic from a certain WEB server within 240 hours; then we use the traffic replay toolkit to reproduce the saved traffic file; at the same time, the mining model in the cluster captures and mines traffic data, so that the stream data-parallel mining model can mine WEB traffic data.

Under normal circumstances, the traffic of the WEB server should be IP packets following HTTP, HTTPS and other protocols. If there is a network attack or intrusion (such as DDOS and TearDrop, etc.), the network traffic data in a certain period of time has certain abnormal characteristics, such as the statistical count of the source address in the seven-tuple. This abnormal characteristic is normal. There have a big difference in flow.

Due to the small traffic of the WEB server (KB-MB level), the simulation of a large traffic and large concurrent environment cannot be completed, so that the mining algorithm

does not consume all the resources that can be allocated. Therefore, in specific processing, by collecting network traffic data within 240 hours, and replaying the traffic to the network within a few hours, considering the carrying capacity of the cluster, the memory usage of the mining model in large-traffic and large-concurrency environment is usually relatively large. In order to avoid the emergence of a deadlock, the maximum memory that can be allocated by each scheduling strategy is limited to 20GB.

Fig. 4 records the value of the mining throughput rate under the Capacity strategy, Fair strategy and RepRM strategy scheduling of the parallel mining model at different flow rates. Fig. 5 records the value of the memory usage. In Fig. 4, when the network data flow rate is low, the Capacity strategy, the Fair strategy and the RepRM strategy can effectively mine the data. With the increase of network data traffic, the throughput rates of the three resource scheduling strategies have gradually increased. When the maximum allocable memory condition of 20G memory is reached, the maximum throughput rate of the Capacity strategy, the Fair strategy and the RepRM strategy is about 3Gbps, 4Gbps and 6Gbps, respectively.

Fig. 5 shows the memory occupancy of the parallel mining model under the Capacity strategy, Fair strategy and RepRM strategy scheduling under different network data flow rates. As shown in Fig. 5, when the network data flow rate is low, the memory footprint of the three resource scheduling strategies is roughly the same, and the value is relatively low. At the same time, it can be seen that the Capacity strategy and the Fair strategy are similar, and slightly larger than the memory occupied by the RepRM strategy. In addition, with the increase of traffic, the memory consumption of the three strategies has gradually increased. When the flow rate reaches 4Gbps, the Capacity strategy and the Fair strategy have exhausted 20G of memory, and when the flow rate reaches 7Gbps, the RepRM strategy has exhausted 20G of memory.

The RepRM resource scheduling strategy considers the location of the data copy in the memory allocation, which can better eliminate the network transmission time of the data copy, so that the mining model can effectively mine. Based on the experimental results in Fig. 4 and Fig. 5, compared to the Capacity strategy and the Fair strategy, RepRM's resource utilization for parallel data mining has increased by about 40%.

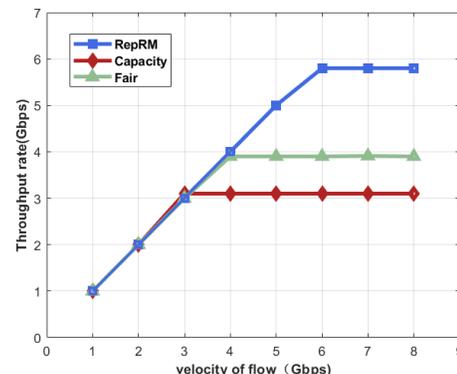


Fig. 4. The Effect of Flow Rate on Throughput.

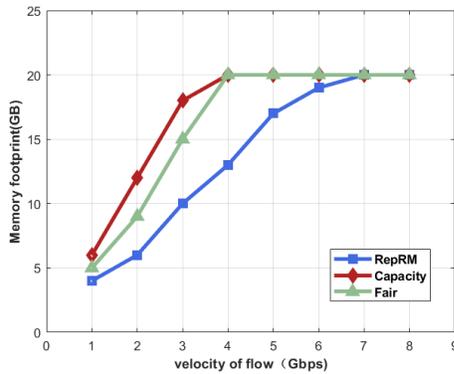


Fig. 5. The Effect of Flow Rate on Memory Consumption.

Table VII shows the accuracy and window size of parallel mining when using the RepRM strategy to schedule resources under different traffic conditions. According to Table VII, when the traffic reaches 6Gbps, because the RepRM resource scheduling strategy has exhausted 20GB of available memory, the network data packet mining capability of the mining model cannot be increased anymore. After that the sampling frequency of the window can only be modified. Mining is carried out, but the sampling method of streaming data seriously affects the accuracy of mining, which makes the accuracy rate continue to decline.

TABLE VII. PARALLEL MINING EXPERIMENTS USING REPRM RESOURCE SCHEDULING

Model	DELL FX2(Including 4 blade servers)
Master	128GBRAM/ 2TBDisk/Operating System CentOS 6.5
Slave1-Slave3	64GBRAM/ 2TBDisk/Operating System CentOS 6.5
Hadoop version	Cloudera 2.2
Network environment	1000M

V. CONCLUSION AND FUTURE WORK

For the mining of massive streams of data, resource allocation has always been a hot research topic. Researchers use various models to schedule CPU, memory, bandwidth, etc., in order to achieve ideal mining results. Especially when performing distributed mining of massive flow data, reasonable resource scheduling can achieve better mining results. However, most researchers did not consider the impact of the location of the data copy on the mining effect, and did not use the data copy as a resource for scheduling.

According to this, for distributed data mining, this paper takes the data copy in the cloud platform as a resource and incorporates it in the resource scheduling strategy for consideration, and realizes a copy-aware resource scheduling strategy RepRM. The RepRM strategy uses data copies as data resources, memory as computing resources, and uses a dynamic programming method to uniformly schedule data resources and computing resources, to improve the adaptability of the mining model to the data and computing resources in the cloud platform. In RepRM, data copies are regarded as resources that need to be allocated. At the same time, in order

to solve the problem of data copy ratio, this paper adopts the dynamic programming method to achieve the maximum mining throughput of the cluster.

Then, this paper conducts simulation tests on parallel mining of streaming data through experiments. The test results prove that the RepRM resource scheduling strategy proposed in this paper has obviously advantages compared to the original resource scheduling strategy of Hadoop itself. After the homogeneous cluster is scheduled through the RepRM strategy, memory resources are saved by about 40-50% during parallel mining of streaming data, and the throughput rate is increased by 20%-30%. After the heterogeneous cluster is scheduled through dynamic planning, the throughput of parallel mining of streaming data is increased by 30-40%, saving about 40% of memory resources.

We aim to extend the RepEM strategy to the use of heterogeneous clusters by taking both the data copy and memory resource as resources and use Lattice point method to solve the problem. What's more, we plan to add more kinds of datasets for further research.

REFERENCES

- [1] Li Qiao, Zheng Xiao. Research Survey of Cloud Computing. Computer Science 2011, 38, 32-37.
- [2] Chi Xuebin, Gu Beibei, Wu Hong, et al. Analysis of the development status of high-performance computer systems and platforms[J]. Computer Engineering and Science, 2013, 35(11): 6-13.
- [3] Barddal J P, Bifet A. A Survey on Ensemble Learning for Data Stream Classification. Acm Computing Surveys 2017, 50, 23. 10.1145/3054925
- [4] C. Franke. Adaptivity in Data Stream Mining[D]. University of California at Davis, 2009.
- [5] Gomes H M, Bifet A, Read J, et al. Adaptive random forests for evolving data stream classification. Machine Learning 2017, 1-27. 10.1007/s10994-017-5642-8.
- [6] Wael Khallouli, Jingwei Huang. Cluster resource scheduling in cloud computing: literature review and research challenges. The Journal of supercomputing 2021. 10.1007/s11227-021-04138-z.
- [7] Ghodsi A, Zaharia M, Hindman B, Konwinski A, Shenker S, Stoica I. Dominant resource fairness: fair allocation of multiple resource types. Proceedings of the 8th USENIX Conference on Networked Systems Design and Implementation 2011, pp 323-336. 10.5555/1972457.1972490.
- [8] Wang W, Li B, Liang B. Dominant resource fairness in cloud computing systems with heterogeneous servers. INFOCOM, 2014 Proceedings IEEE. IEEE 2014, pp 583-591. 10.1109/INFOCOM.2014.6847983.
- [9] Khamse-Ashari J, Lambadaris I, Kesidis G, Urgaonkar B, Zhao Y. Per-server dominant-share fairness (ps-dsf): a multi-resource fair allocation mechanism for heterogeneous servers. 2017 IEEE International Conference on Communications (ICC). IEEE 2017, pp 1-7. 10.5555/1972457.1972490.
- [10] Niu Z, Tang S, He B. Gemini: An adaptive performance-fairness scheduler for data-intensive cluster computing. 2015 IEEE 7th International Conference on Cloud Computing Technology and Science (CloudCom). IEEE 2015, pp 66-73. 10.1109/CloudCom.2015.52.
- [11] Shao Y, Li C, Gu J, Zhang J, Luo Y. Efficient jobs scheduling approach for big data applications. Comput Indus Eng 117 2018. 249-261;10.1016/j.cie.2018.02.006.
- [12] Ali Shakarami, Mostafa Ghobaei-Arani, Ali Shahidinejad, Mohammad Masdari, Hamid Shakarami. Data replication schemes in cloud computing: a survey. Cluster Computing 2021, pp 2545-2579. 10.1007/s10586-021-03283-7.
- [13] Najme Mansouri, Mohammad Masoud Javidi. A review of data replication based on meta-heuristics approach in cloud computing and data grid. Soft Computing 2020, pp 14503-14530. 10.1007/s00500-020-04802-1.

- [14] Cui L, Zhang J, Yue L, Shi Y, Li H, Yuan D. A genetic algorithm based data replica placement strategy for scientific applications in clouds. *IEEE Trans Serv Comput* 2018, 11(4):727–739. 10.1109/TSC.2015.2481421.
- [15] Chunlin L, Ping WY, Hengliang T, Youlong L. Dynamic multi-objective optimized replica placement and migration strategies for SaaS applications in edge cloud. *Future Gener Comput Syst* 2019, 100:921–937;10.1016/j.future.2019.05.003.
- [16] Huang X, Wu F. A cost-effective data replica placement strategy based on hybrid genetic algorithm for cloud services. *International conference on research and practical issues of enterprise information systems* 2018, pp 43-56. 10.1007/978-3-319-99040-8_4.
- [17] Khojand M, Fatan Serj M, Ashrafi S, Namaki V. Predicting dynamic replication based on fuzzy system in data grid 2018. arXiv:1804.02963
- [18] Salem R, Salam MA, Abdelkader H, Awad A, Arafa A, An artificial bee colony algorithm for data replication optimization in cloud environments. *IEEE Access* 2019, 7:1–12. 10.1109/ACCESS.2019.2957436.
- [19] Christiansen H. A survey of adaptable grammars. *Sigplan Notices* 1990, 25,35-44. 10.1145/101356.101357.
- [20] Li Zhilin, Ou Yigui. *Mathematical modeling and analysis of typical cases*. Beijing: Chemical Industry Press, 2007.10.3969/j.issn.1000-4076.2006.05.035.
- [21] Ming, Sun Lingyu, Zhu Ping. Research on the Load Balancing Task Scheduling Algorithm Based on Cellular Automata in Cloud Computing. *Small Microcomputer System* 2016, 37, 2212-2216.

Use of Neural Networks in the Adaptive Testing System

Ekaterina Vitalevna Chumakova¹, Tatiana Alexandrovna Chernova², Yulia Aleksandrovna Belyaeva³
Dmitry Gennadievich Korneev⁴, Mikhail Samuilovich Gasparian⁵
Moscow Aviation Institute, Moscow, Russia^{1, 2, 3}
Plekhanov Russian Economic University, Moscow, Russia^{4, 5}

Abstract—The paper examines the issues of the use of adaptive testing systems in terms of their incorporation in artificial neural network modules designed to solve the problem of choosing the next question, thereby forming an individual testing trajectory. The study presents an analysis of data affecting the quality of problem-solving, proposes a general modular structure of a system, and describes the main data flows at the input of an artificial neural network. The solution proposed for the problem of choosing the difficulty of the question is to use feedforward neural networks. Different architectures and parameters of training artificial neural networks (weight update mechanisms, loss functions, the number of training epochs, batch sizes) are compared. As an alternative, the option of using recurrent long-short term memory networks is considered.

Keywords—Adaptive testing system; artificial neural network; machine learning

I. INTRODUCTION

Adaptive testing is a technology of determining the level of knowledge of the tested subject, in which each upcoming question is automatically selected based on previous answers. The advantage of such testing, as seen by specialists, is the opportunity to determine the testee's level of knowledge more comprehensively and accurately. The problems of developing adaptive tests are topical not only as part of testing students, for example, for the purpose of developing individual learning trajectories, but also in other spheres that require assessment of the subject's competencies and personal intellectual and psychophysiological characteristics. Increasing interest in adaptive testing is demonstrated, for instance, by HR specialists in large companies concerned with recruiting new specialists and testing current employees.

At the heart of adaptive testing systems is an intelligent way to select questions individually for each test subject, based on the answers in all previous steps of the test. The degree of adaptation of the test depends on the number of parameters considered, such as the level of complexity and the number of tasks that are proposed to be completed [1]. Of the greatest interest for research are flexible adaptive systems that allow one to achieve a large variability of tests with high accuracy and reliability in determining the level of training and make the testing process itself look like an oral exam with a teacher. Therefore, the purpose of the study is to organize the process of intellectual choice of a topic (a thematic block of questions) and determine the complexity of the next

question, considering previous answers and the complexity of previously asked questions, as well as the connectivity of topics (blocks) and response time as a factor in guessing or searching for an answer in part of the adaptive testing system.

With the advancement of smart technologies, the development of new methods and the resolution of particular problems using computerized adaptive testing (CAT) is attracting ever-increasing interest from specialists.

At present, we can note three major directions of research, along which CAT methods are being developed:

- Item Response Theory (IRT). IRT is a set of related psychometric theories that serves as a foundation for subject assessment. At the basis of this theory lie mathematical models and logical functions characterizing the relationship between the subject's features (characteristics, knowledge) and the probability of correct answers.
- Bayesian Belief Network (BBN). BBN is a formal graphical language for representing and conveying decision scenarios that require reasoning under uncertainty.
- Artificial Neural Networks (ANNs). ANN is an information processing paradigm based on mechanisms similar to the operation of neurons in the human brain. An ANN is comprised of a certain number of interconnected nodes (neurons) that process information and transmit signals to other neurons based on the results of processing.

It can be stated that the majority of current theoretical and applied research in CAT concerns primarily the use of ANNs. In the field of testing, ANNs are most often proposed to be used as the final module for scoring. In several works [1-3], there were attempts to solve the problem of intellectual choice of a question in the form of determining the level of difficulty of the next question based on one previous one based on the correct or incorrect answer.

Currently, there are many different frameworks for creating ANNs, which makes this mechanism more accessible. However, creating and training a neural network that could offer a substantial advantage over traditional testing requires advanced theoretical knowledge and a fair number of experiments.

At present, researchers do not have a universal approach to creating neural networks for CAT. There arise questions related to the choice of the type and architecture of the network, as well as the number of training examples needed for an acceptable quality of recommendations generated by the network. Contemporary studies only offer general recommendations on these issues. For instance, Golovko and Krasnoproshin [2] suggest quite a large interval of the number of hidden layers.

It is also worth noting the lack of comprehensive works which, having an in-depth look at the entire process of creating a neural network, also describe the applied network learning technologies, which are an important element in ensuring the performance of an ANN.

Approaches to choosing the type of network also vary. Researchers employ both the "classical" feedforward neural networks, in which the signal goes sequentially from layer to layer [3, 4, 5], and recurrent neural networks, in which there is feedback between neurons, and the output signal can be transmitted to the input of the neurons of the previous layer [6]. What should be noted as interesting ideas published in a number of papers is the application of the methods of open systems in the creation of neural networks, in particular, the creation of ANN according to the modular principle [6].

II. METHODS

In the course of the study, the process of conducting a knowledge test in technical disciplines was analyzed from the point of view of its general organization and the identification of indicators that affect the course of testing. As disciplines, such disciplines as "Databases", "Informatics", and "Computer Graphics", read in different educational institutions for students of different courses and specialties, were chosen, and four teachers acted as experts.

At the initial stage, a classification of the tasks solved by the ANN within the process was carried out. Studies were carried out on the use of various types of networks for solving CAT problems, on the basis of which the types of network architectures for further research were determined. The place of the ANN in the overall structure of the adaptive testing system was determined. As a result, a general modular structure of the system was proposed and the main data flows entering the ANN input were described. To achieve the universality of the approach, i.e. regardless of the subject of the test, the choice of the next question was proposed to be carried out in two stages using two ANNs: to select a topic and select the level of difficulty of a question for a particular topic.

At the next stage, to solve the problem of determining the complexity of the question, a feed-forward network was considered. A comparison was made of various ANN architectures and training parameters (weight update algorithms, loss functions, number of training epochs, packet sizes).

In accordance with the results obtained at the previous stage, it was proposed to use six input neurons, which are fed with normalized average values of the correctness of answers to questions and their complexity, the number of questions

already asked, the average time of deviation from the expected response time, the assessment of the answer to the last question asked, and its complexity. At the output of the network there are five output neurons corresponding to the difficulty levels of the questions. Thus, for the simplest feed-forward network with one hidden layer of neurons, the 6-m-5 architecture was chosen and, in accordance with general heuristic recommendations, experiments were carried out for $m = 9, 12, 15, 18, 21$.

All results were obtained using the high-level Keras library, which allows you to quickly start at the initial stages of research and get the first results. SGD, Adam, NAdam and RMSprop implemented in Keras were compared as optimizers to achieve faster convergence. The loss function MSE (mean square error) was used together with the optimizer. Training was carried out on a training set of 1500 sets, which accounted for 80% of the general sample, 10% each were validation and test samples prepared by experts. Traditionally, training was carried out for a large number of epochs (50, 100, 20, 350, and 500), experimentally obtained graphs of accuracy versus the number of epochs for a different number of neurons in the hidden layer to determine the most appropriate architecture. The resulting graphs were constructed using cubic spline interpolation.

In order to determine the effect of the data packet size on the learning process for the 6-12-5 architecture, several experiments were carried out with packets of various sizes (5, 10, 15, 20, 25, 30, 50 and 100) and the optimal size for this task was determined.

Similar training experiments on the same general sample were carried out when switching to network architectures that included two and three hidden layers within 9-21 neurons.

At the final stage, as an alternative, the possibility of using a recurrent ANN LSTM (Long-Short Term Memory) network was considered. In accordance with the feature of this type of network, the number of input neurons was reduced to four, to which the question number, answer score, question complexity, and temporal deviation from the normal were applied. For training, the same tools and approaches were used as in training the feedforward network.

III. MODULAR STRUCTURE OF THE ADAPTIVE TESTING SYSTEM

The main objective solved by an adaptive testing system is the identification of a reliable "profile" of the examinee's knowledge in a particular area. In this case, adaptivity is understood not only as the intellectual selection of questions depending on the level of knowledge demonstrated by the subject, but also the extensibility and universality of the system as a whole [7, 8]. It is thereby clear that such a system has to be constructed by the modular principle, which will ultimately give the structure of the system greater flexibility and versatility.

Of particular interest is the smart selection of the next question. The approaches used in practice differ: assigning the subject an equal number of questions on all topics with different levels of complexity; giving the subject more questions on the topics they made mistakes in; or selecting the

questions using a clear preset algorithm [9-13]. In this case, these methods are difficult to regard as smart approaches.

In this paper, by the subject's profile of knowledge in the given area, we understand the level of mastery over the material in each considered topic. The level of mastery has to be determined in accordance with the difficulty of the assigned questions and the accuracy of the given answers. Therefore, all questions from the bank should not only belong to a certain topic but be characterized by a specific level of complexity.

Obtaining a reliable profile of knowledge through testing presupposes selecting the topic and complexity of the question to be assigned next, meaning that the question is selected for a specific subject in view of the number of questions assigned by topic, their association with one another, level of complexity, and the accuracy of answers given at the previous stages. This task is what an ANN is intended to solve.

The choice of the group of the next question is influenced by the following parameters, which serve as input data for the ANN:

- The topics of the question assigned previously;
- The number of the assigned questions (in each topic);
- The difficulty of the assigned questions (in each topic);
- The accuracy (grading) of the given answers (by the topics considering the levels of complexity);
- The relatedness of the topics to each other;
- Response time to the questions assigned previously.

Let us describe the form in which these data can be stored in the system. It is proposed to store a vector (an array with

the dimensionality equal to the number of questions assigned) of structures for each test taker containing:

- Question – topic, number, difficulty;
- Response time, or rather its positive or negative deviation from the expected norm, i.e. the time sufficient to read the question and give a meaningful answer;
- Grade – the degree of response accuracy (1 – correct, 0 – incorrect).

The relatedness of topics is set via the matrix $M[N, N]$ of coefficients varying in the range from 0 to 1, where 0 – the topics are completely unrelated, 1 – related to the highest degree. The coefficient at the intersection of the i -th row and the j -th column shows how related the i -th and the j -th topics are. The matrix is symmetric, with ones on the main diagonal, so the only really significant input values are $(N^2 - N)/2$.

With a large number of test questions (30-50), the number of input parameters is not only large but also constantly changes as new answers are given (growing in an arithmetic progression). This creates major challenges in determining the architecture of an ANN, as well as complicates the preparation of training and test data sets and the training process itself. In addition, the number of input parameters also changes when so does the number of questions in the test, which eliminates all possible universality.

To resolve this issue, we use two ANNs, one determining the topic of the next question and the other setting the difficulty level. To preserve the number of input parameters and make it constant, we integrate the ANNs with each other by means of algorithmic modules, which perform preliminary mathematical preparation of input values. The resulting general structure of the testing system, which has a hybrid architecture, is presented in Fig. 1.

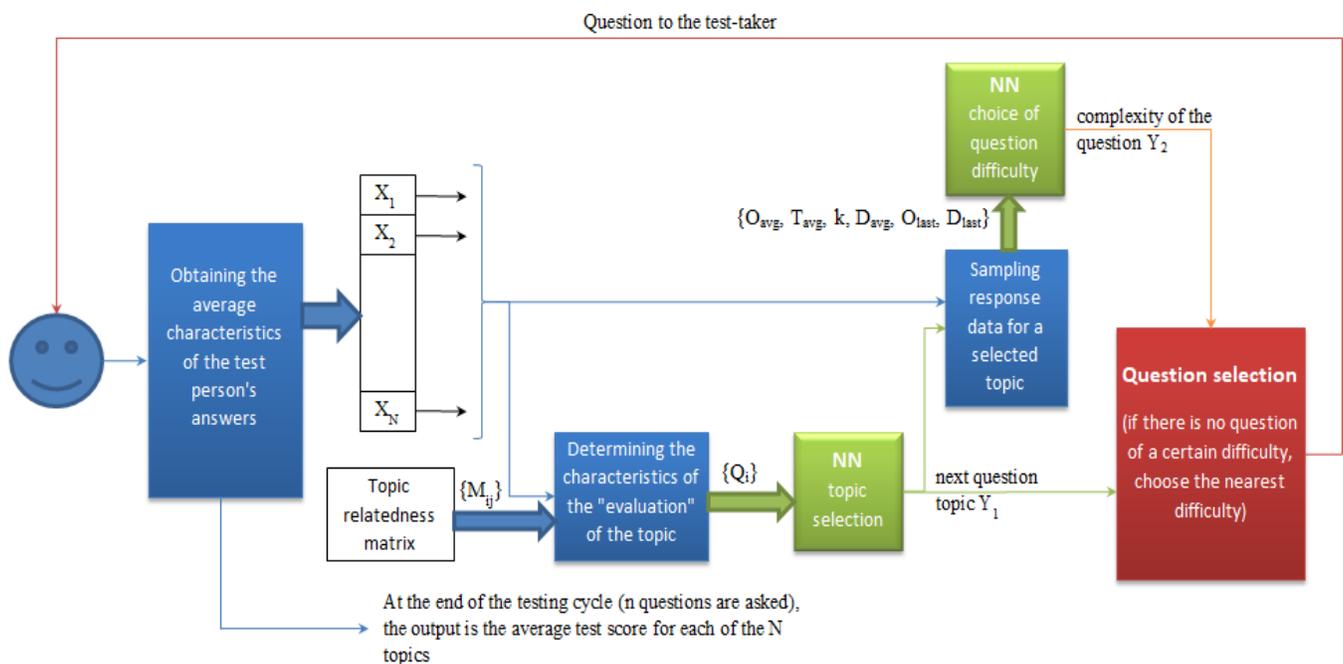


Fig. 1. General Scheme of Functioning of the Adaptive Testing System.

In the case of a feedforward neural network, consideration of all answers given to all of the questions is proposed to be performed by inputting in the networks the average parameters of answers for each topic $\{X_i\}$, for which purpose a respective module is included in the system. In the case of a recurrent ANN, the aforementioned module can be absent due to the capacity of the network itself to account for its previous states. In both cases, the number of inputs with a large number of topics is reduced insignificantly, but stays constant at all stages of testing for any number of questions. The efficiency of using particular types of ANNs is suggested to be assessed in further research.

The choice of the topic in the process of creating the individual testing trajectory needs to take into consideration both the subject's answers and the relatedness of the topics in order to, first, test proficiency in the material across different topics and, second, optimize the total number of questions asked in each topic. At the input of the topic selection ANN, it is proposed to put a vector of "assessment" coefficients of the topics $\{Q_i\}$, which are essentially derived by summing up the share of each answer minding the level of difficulty and relatedness of the topics. Mathematically, this can be expressed as the product of the topic connectivity matrix described earlier and the vector of averaged difficulty-weighted answers.

Once the topic is selected, the second ANN module should determine the difficulty of the upcoming question based on the averaged data for the already specified topic. The logic of decreasing or increasing the difficulty of questions is laid down during training based on the training sets provided by experts according to the given requirements.

As a result, the test taker receives a question selected from the question bank, the difficulty and topic of which are individually selected depending on the test taker's previous answers.

IV. NEURAL NETWORK MODULE FOR DETERMINING THE LEVEL OF DIFFICULTY OF A QUESTION

At the first stage, we design and train an ANN module responsible for selecting the difficulty level. Specification of the difficulty of the question after its topic is already selected largely decreases the number of input parameters that can affect the choice of the question. In this case, it is necessary to consider the parameters of the subject's answers only in one specific topic. For this very reason, this ANN module is preceded by the block of selection of the test taker's performance on a specific subject topic.

In the course of the study, the feasibility of using various ANN models was analyzed from the point of the correspondence of this task to the specific class of tasks solved by particular types of ANNs. This task, however, cannot be unequivocally attributed to a single type since, on the one hand, determining the difficulty of the next question is a classification task, i.e. determining the class of difficulty based on the subject's performance, while on the other, this task involves predicting the real level of knowledge based on the previous answers.

Although there are no specific architectures designed to solve classification tasks, the most commonly used type, in this case, is a multilayer feedforward neural network. For tasks based on sequences, a special type of ANN – a recurrent network – is used. It is impossible to determine in advance exactly which of the architectures is best suited for the task. Therefore, we focus on a more detailed study of two variants of networks, specifically:

- A feedforward neural network, in which all layers are connected with one another directly and sequentially – without feedback or delay lines.
- A recurrent Long-Short Term Memory (LSTM) network, which receives information from the previous passes, thereby being capable of learning long-term dependencies.

The well-known and obvious disadvantage of the latter is their high demands for hardware and resources, both in training (the training process takes a significant amount of time) and in startup.

Next, we will more closely consider the option of using a feedforward neural network. To account for all the previous answers on the topic, the number of which at a certain stage can be random, the input fed to the ANN should include the average values of the accuracy of the given answers and the difficulty of the questions, the number of questions already asked, as well as the average deviation from the expected answer time as a kind of indicator of guessing or searching for answers. In addition to the average values, which do not provide complete information for decision-making in this task even for a human, the network input also includes the mark for the last question answered and its complexity.

Thus, the input layer contains 6 neurons, the output layer – depending on the number of question difficulty levels. In our case, there are 5 neurons, which are aggregated into the last layer containing one neuron. For the final determination of the network architecture, it is necessary to determine the number of hidden layers and the number of neurons in them.

Following the recommendations of Golovko and Krasnoproshin [2], for a network with n-m-p architecture and training sample volume L, the number of neurons in the hidden layer should meet the following condition:

$$\log_2(p) < m < (L - p) / (n + p + 1) \quad (1)$$

Herein, the upper bound is derived from the condition that the training sample size exceeds the number of adjustable parameters. At the same time, there are heuristic rules according to which the size of the training sample must, at least, by an order of magnitude, exceed the number of adjustable parameters to obtain an error of 10%, and the number of hidden layer neurons must, at least, exceed the size of the input by 1.5-2 times [14].

If the use of a perceptron with one hidden layer fails to provide the required accuracy and generalizability of the network, then a neural network with more than one hidden layer is used. The optimal network architecture is also determined via genetic and evolutionary algorithms, which

have their own practical features and limitations. Therefore, at the first stage, we examine a network with architecture n-m-p, for which m and L should satisfy the conditions described earlier.

Since there are no exact methods for estimating the complexity of the problem to be solved and the learning algorithm, which are the determining factors for choosing the amount of data in machine learning, the sufficient amount of data cannot be determined in advance. On the basis of the above recommendations, it is possible to roughly estimate the required size of the general population of raw data and the number of neurons in a hidden layer. We will proceed from the fact that the minimum number of neurons in a hidden layer should be $m=9-12$, suggesting that the volume of the training sample should be, on the one hand, $L>110-150$, and on the other hand, $L\geq 1150-1500$, i.e. an order higher than the number of adjustable parameters for the given m. In the absence of other points of reference, the size of the training sample is set to be 1500 items. According to the generally accepted ratios, the training sample must be 80% of the general population, the verification (validation) sample – 10% (150 items), and the test (control) sample – 10% (150 observations). For the chosen training set, it is advisable to study a number of architectures for $m = 9, 12, 15, 18, 21$ neurons.

The requirements taken into account when preparing the samples are that they should contain a sufficient number of unique examples, should not contain duplicates, contradictions, omissions, and anomalous values, and that the numerical ratio of objects of different classes in each sample should be the same as in the initial general population [15].

In particular, the data structure is affected by the method of network learning. In our case, the “teacher – student” method is employed, as this method is the most commonly used for classification and prediction tasks. Training without a teacher is used in statistical and language models, as well as, for example, in the tasks of clustering and data compression, which does not correspond to the conditions of our task.

Two libraries (frameworks) are considered for the ANN modeling – Keras and PyTorch. These libraries differ in the levels of API and the ways of describing and running network training. Nevertheless, they produced similar results for the described network architecture and training on the same training sets. All the presented results are obtained with the Keras library, which makes it possible to easily create networks and simplify testing of training models, offering additional convenience in the initial stages of research and obtaining first results.

The quality of training of the model is determined not only by its structure and training set but a number of training parameters: the weight update algorithm (optimizer), the loss function, the number of training epochs, and batch size. Below we compare the most popular optimization methods, SGD, Adam, NAdam, and RMSprop, implemented in Keras to achieve faster convergence. The analysis shows that for the problem under study, the best results in terms of accuracy are demonstrated by Adam. The MSE loss function (root-mean-square error) is used for all optimizers.

Traditionally, training is performed for a large number of epochs, which is usually determined experimentally and is sufficient to obtain minimal error and high accuracy [16-18]. In this study, the training of networks with different numbers of hidden layer neurons is performed in the span of 50, 100, 200, 350, and 500 epochs. The results reveal dependencies of accuracy on the number of training epochs for networks with different numbers of neurons in the hidden layer, which are presented graphically below (Fig. 2).

The last noted parameter is the batch size (`batch_size`), i.e. the number of examples in the sample run through the network after which the weight coefficients are updated. Keras implements mini-batch gradient descent with the recommended batch size being 32. Meanwhile, the generalization ability can decline not only when the batch size, which is chosen experimentally, is reduced, but also when it is increased, which is due to the inner noise in the gradient estimation [19]. Several experiments with different batch sizes (5, 10, 15, 20, 25, 30, 50, and 100) are conducted for the 6-12-5 network, resulting in the best generalization ability obtained with a batch size of 50.

The network with the 6-15-5 architecture shows the best accuracy, the obtained learning curves are shown in the graph below (Fig. 3).

An increase of the number of neurons in the hidden layer, the number of hidden layers, and training epochs, as well as mixing the data and change of the learning rate, by means of Keras does not result in an increase in network performance, accuracy rates remain at an average level of 83-85%. The effect of retraining is also not observed. The conclusion from the conducted experiments is that to further improve the accuracy, the general sample of examples needs to be analyzed in terms of the completeness and complexity of the model.

Next, we examine a recurrent network with a similar number of neurons per layer [20, 21]. Proceeding from the fact that an LSTM network processes the temporal sequence of input data while preserving the internal state obtained when processing the previous items, it is not necessary to calculate averaged values to account for all previously received responses. The number of network inputs can be reduced to 4: question number, answer score, question difficulty, and response time deviation from normal. In general, the set of input parameters of the LSTM network will not differ from the previously considered case for the feedforward neural network.

The exception is the form in which training and test sets are presented, each of which, in fact, is a sequence of answers to the questions. In this case, the optimization methods and the functions of activation and evaluation of training results used for the LSTM network will be the same as for the feedforward neural network.

A number of experiments yield similar results (model accuracy of 95% and accuracy on the test sample of 80%) and learning curves, one of which is shown below (Fig. 4). The retraining effect is observed already at 100-150 epochs of training, indicating that the network remembers all the examples and its training requires large samples.

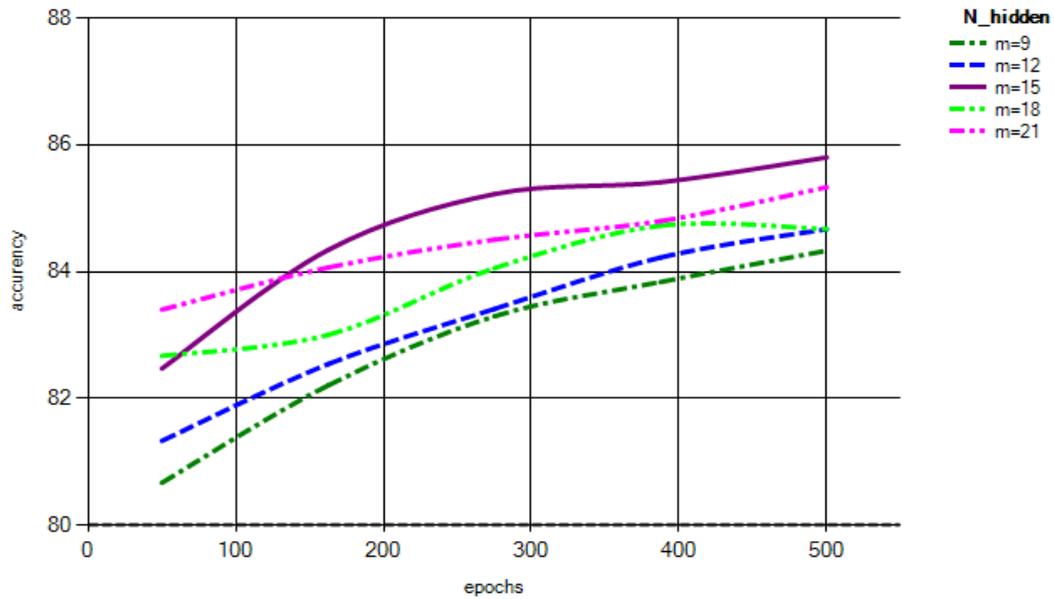


Fig. 2. Dependencies of Accuracy on the Number of Training Epochs for different Numbers of Neurons in the Hidden Layer.

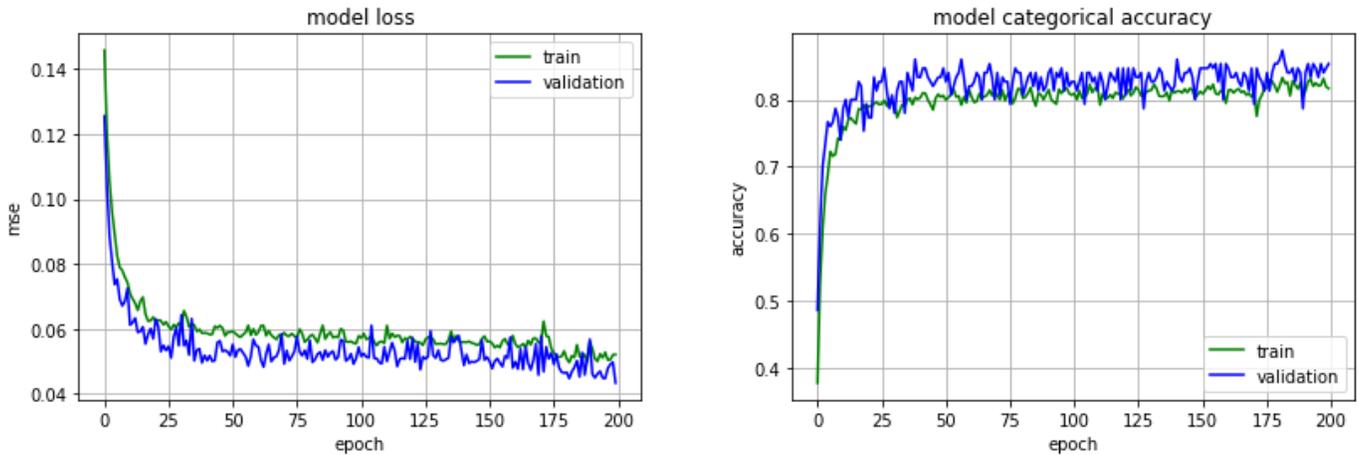


Fig. 3. Learning Curves for Networks with 6-15-5 Architecture.

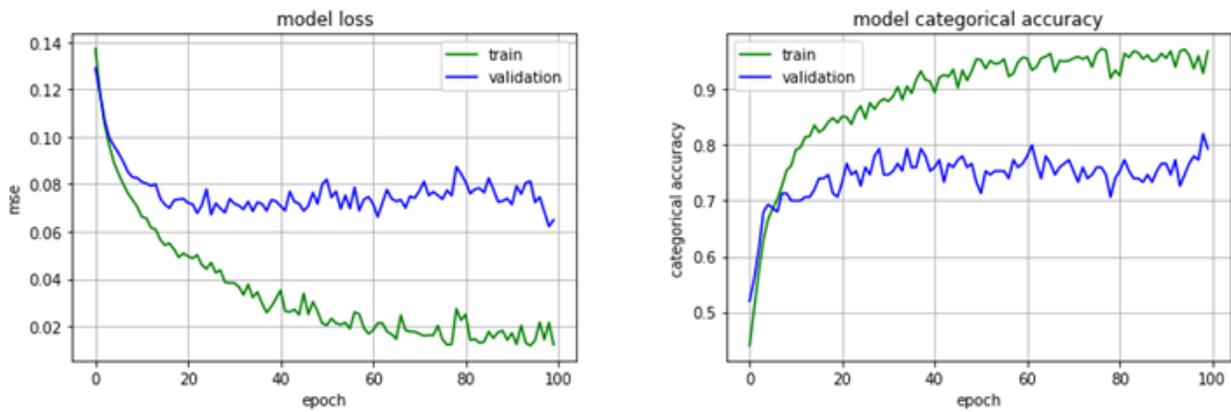


Fig. 4. LSTM Network Learning Curve.

It is not necessary to talk about the advantage of one of the types of networks (forward propagation networks and LSTM networks) for solving the problem of determining the next question in adaptive testing problems. Both types of networks require more detailed preparation of the training set. For direct propagation networks, this is primarily due to the complexity of the model, namely, the nature of the network input data – the averaged values of all previous answers. In this case, to complete the description of the model, more unique and consistent training and test cases are required. The form of the input data of the recurrent network is simpler, but the task of determining the complexity of the next answer based on all the previous ones remains, and, therefore, the training data must represent different testing trajectories at different stages. In addition, the number of adjustable parameters is several times greater than that of the direct distribution network, which also increases the requirements for the volume and quality of the training sample. Preliminary analysis of the training set revealed the presence of gaps in certain data ranges, as well as a certain uneven representation, which, if randomly mixed, can lead to learning failures.

V. DISCUSSION

In general, based on the study, we can conclude that the obtained accuracy of the direct propagation network of 83-85% is quite sufficient for its use in the adaptive testing system. A well-known and obvious disadvantage of using LSTMs is their demands on equipment and resources, both during training (the training process takes a significant time) and during startup, in our case, it is supplemented by increased requirements for the training set and, despite the obtained accuracy of 80%, puts questioning the expediency of further study of LSTM networks in solving this problem.

Using this solution to determine the level of complexity of the next question will consider all the answers given to the subjects, the levels of complexity of the questions, as well as the time of answers within each thematic block, and in the future and considering their thematic connection, in contrast to the previously proposed solutions, in which, only one previous test step is considered, and the accuracy for the recurrent network is 75%.

VI. CONCLUSION

The paper proposes a new structure for organizing an adaptive testing system based on the use of neural network modules, training of the neural network responsible for determining the level of complexity of the question asked is performed. The results obtained can be used in the construction of the second module of the ANN of the system, which is responsible for choosing a topic.

In addition to selecting the topic of the next question, an ANN can be entrusted with the task of moving to the next test stage (assigning the next question), i.e. deriving a reliable knowledge profile with an individual long test trajectory. The study of this possibility is interesting in terms of how optimal the number of given questions will be, whether the system will not go into infinite test mode, or the tests will be too short.

In addition, it is necessary to answer the question of the need to improve the efficiency of an already implemented

network, and, consequently, to conduct research on methods to improve the efficiency of networks, including finer tuning of parameters and learning algorithms, as well as architecture.

In general, the introduction of the proposed tools will allow organizing the process of adaptive testing, with an intelligent selection of questions depending on the demonstrated level of knowledge of the test person to form an individual testing trajectory in order to determine the reliable level of knowledge of the test subject for the optimal number of questions asked.

REFERENCES

- [1] R. P. Chalmers, "Generating Adaptive and Non-Adaptive Test Interfaces for Multidimensional Item Response Theory Applications", *Journal of Statistical Software*, vol. 71(5), 2016, pp. 1–38. <https://doi.org/10.18637/jss.v071.i05>.
- [2] V. A. Golovko, and V. V. Krasnoprosin, "Neural Network Technologies of Data Processing" [Neurosetevye tekhnologii obrabotki dannikh], Minsk: BSU, 2017.
- [3] E. Iu. Savchenko, "Application of modified neural network learning algorithms for adaptive testing tasks" ["Primenenie modifitsirovannykh algoritmov obucheniia neironnykh setei v zadachakh adaptivnogo testirovaniia"], *Nauchnyi aspekt*, no. 4, 2012. <https://na-journal.ru/4-2012-tehnicheskie-nauki/159-primenenie-modifitsirovannykh-algoritmov-obucheniia-neironnykh-setej-v-zadachah-adaptivnogo-testirovaniia>.
- [4] A. Petrovskaya, D. Pavlenko, K. Feofanov, V. Klimov, "Computerization of learning management process as a means of improving the quality of the educational process and student motivation", *Procedia Computer Science*, vol. 169, 2020, pp. 656-661 <https://doi.org/10.1016/j.procs.2020.02.194>.
- [5] A. Grigorev, and V. Mamaev, "On the application of neural networks in knowledge testing" ["O primeneniia neironnykh setei v testirovaniia znaniia"], *Nauchnoe Priborostroenie*, vol. 4, no 26, pp. 77–84, 2016.
- [6] S. Sh. M. Jafri, "Computerized adaptive testing using neural networks", 2007. https://www.researchgate.net/publication/228720628_Computerized_adaptive_testing_using_neural_networks.
- [7] O. Iu. Nikiforov, "Use of adaptive computer-based testing systems" ["Ispolzovanie adaptivnykh sistem kompiuternogo testirovaniia"], *Gumanitarnyye nauchnyye issledovaniia*, vol. 4, 2014. <https://human.snauka.ru/2014/04/6274>.
- [8] D. S. Zhadaev, A. A. Kuzmenko, and V. V. Spasennikov, "Peculiarities of neural network analysis of students' training level in the process of adaptive testing of their professional competencies" ["Osobennosti neirosetevogo analiza urovnia podgotovki studentov v protsesse adaptivnogo testirovaniia ikh professionalnykh kompetentsii"], *Vestnik Bryanskogo gosudarstvennogo tekhnicheskogo universiteta*, vol. 2, no. 75, pp. 90-98, 2019.
- [9] V. A. Pesoshin, V. V. Zvezdin, A. N. Iliukhin, R. R. Saubanov, and R. R. Saubanov, "Automated testing system as a tool to improve the quality of knowledge assessment" ["Avtomatizirovannaia sistema testirovaniia kak instrument povysheniia kachestva otsenki znaniia"], *Mashinostroyeniye i komp'yuternyye tekhnologii*, vol. 7, pp. 137–142, 2016.
- [10] J. Rodríguez-Cuadrado, D. Delgado-Gómez, J.C. Laria, S. Rodríguez-Cuadrado, "Merged Tree-CAT: A fast method for building precise computerized adaptive tests based on decision trees", *Expert Systems with Applications*, vol. 143, 1 April 2020, 113066. <https://doi.org/10.1016/j.eswa.2019.113066>.
- [11] D. Pavlenko, L. Barykin, S. Nemeshev, E. Bezverhny, "Individual approach to knowledge control in learning management system", *Procedia Computer Science*, vol. 169, 2020, pp. 259-263. <https://doi.org/10.1016/j.procs.2020.02.162>.
- [12] H. Özyurt, Ö. Özyurt, A. Baki, B. Güven, "Integrating computerized adaptive testing into UZWEBMAT: Implementation of individualized assessment module in an e-learning system", *Expert Systems with Applications*, vol. 39, Issue 10, August 2012, pp. 9837-9847. <https://doi.org/10.1016/j.eswa.2012.02.168>.

- [13] D. A. Pominov, L.S. Kuravsky, P.N. Dumin, G.A. Yuriev, "Adaptive trainer for preparing students for mathematical exams", *International Journal of Advanced Research in Engineering and Technology*, vol 11, Issue 11, November 2020, pp. 260-268.
- [14] S. Haykin, "Neural networks: a complete course" ["Neironnye seti: polnyi kurs"], Moscow: Viliams, 2019.
- [15] X. Xu, T. Liang, J. Zhu, D. Zheng, T. Sun, "Review of classical dimensionality reduction and sample selection methods for large-scale data processing", *Neurocomputing*, vol. 328, 7 February 2019, pp. 5-15 <https://doi.org/10.1016/j.neucom.2018.02.100>.
- [16] K. M. Ang, W.H. Lim, S.S. Tiang, C.K. Ang, E. Natarajan, M.K.A. Ahamed Khan, "Optimal training of feedforward neural networks using teaching-learning-based optimization with modified learning phases", *Lecture Notes in Electrical Engineering*, vol. 770, 2022. pp. 867-887.
- [17] K. Gorshkova, L. Tugashova, V. Zueva, M. Kuznetsova, "Optimizing deep learning methods in neural network architectures", *International Review of Automatic Control*, vol. 14, Issue 2, 2021, pp. 93-101.
- [18] Y. Zhao, "Research on management model based on deep learning", *Complexity*, vol. 2021, 2021, pp. 9997662.
- [19] N. Shirish Keskar, D. Mudigere, J. Nocedal, M. Smelyanskiy, P. T.P. Tang, "On Large-Batch Training for Deep Learning: Generalization Gap and Sharp Minima", Published as a conference paper at ICLR 2017 <https://arxiv.org/pdf/1609.04836.pdf>.
- [20] Y. Huo, D.F. Wong, L.M. Ni, L.S. Chao, J. Zhang, "Knowledge modeling via contextualized representations for LSTM-based personalized exercise recommendation", *Information Sciences*, vol. 523, June 2020, pp. 266-278. <https://doi.org/10.1016/j.ins.2020.03.014>.
- [21] Md. Kowsher, A. Tahabilder, Md. Z. Sanjid, N. J. Prottasha, Md. S. Uddin, Md A. Hossain, Md. A. K. Jilani, "LSTM-ANN & BiLSTM-ANN: Hybrid deep learning models for enhanced classification accuracy", *Procedia Computer Science*, vol. 193, 2021, pp. 131-140. <https://doi.org/10.1016/j.procs.2021.10.013>.

Development of Ontology-based Domain Knowledge Model for IT Domain in e-Tutor Systems

Ghanim Hussein Ali Ahmed, Jawad Alshboul, László Kovács

University of Miskolc, Faculty of Mechanical Engineering and Informatics, Miskolc, Hungary

Abstract—Ontology as a technology has been studied in many areas and is being used in several fields. A number of studies have utilized ontology to manage problems such as interoperability in teaching materials, modeling and enriching education resources, and personalizing learning content recommendations in the educational context. A possible reason for the lack of success may be that simply posting lecture notes on the internet does not provide enough learning and training. However, this situation can improve by using education software like an e-tutoring system. The e-tutor system has built-in modules to track students' performance and personalize learning according to an adaptation of students' learning styles, knowledge levels, and proper teaching techniques in e-learning systems. e-Tutor is an excellent area in the context of electronic instruction since it provides adequate aid for learners and becomes increasingly important for individual and collaborative learning. Thus, there has been significant interest in adopting e-tutoring to facilitate learning processes and enhance learners' performance. This paper represents a domain knowledge model for an e-tutoring system that enables knowledge to be stored in such a way separated from the domain of interest and assists in storing transfer and prerequisite knowledge relationships. This innovative technique is helpful for the students in improving their learning progress. This paper introduced a domain knowledge model for an e-tutor system to support the way of teaching and learning process. The model implementation is developed in Python and Owlready2. Two types of ontologies are provided: general concepts of the domain knowledge ontology and specific domain knowledge ontology. This solution represents the knowledge to be learned, delivers input to the expert model, and eventually provides specific feedback, selects problems, generates guidance, and supports the student model.

Keywords—e-learning; knowledge model; domain model; e-tutor system; SPARQL

I. INTRODUCTION

E-learning is a technique of learning and teaching utilizing Information and Communication Technology. It means that the growth of online learning is a collaborative task including several researchers in several disciplines such as educational design and learning material [1]. e-Learning system is frequently getting popular in the academic community due to the advantages of using it anywhere, and anytime [2]. However, it tends to be most commonly employed for web-based education to access online courses via the internet [2]. A possible reason for the lack of success may be that merely posting lecture notes on the internet does not provide enough learning and training [2]. This situation can be improved by using education software such as an e-tutoring system. The e-tutoring system includes built-in modules to monitor a learner's

performance and personalize education based on adaptation to learners' learning style, existing knowledge level, and suitable teaching approaches in e-learning systems. The e-tutoring framework usually applies to computer-based instruction. However, an e-tutoring system generally uses adaptive mechanisms and concepts to solve the current e-learning problems by applying artificial intelligence techniques.

Artificial Intelligence in Education (AIED) society frequently identifies the value of producing technologies with a global reach [3]. e-Tutor system is a computer application that uses artificial intelligence approaches to improve and personalize the teaching and learning processes. e-Tutoring differs from other e-learning systems because it uses a knowledge base to guide the pedagogical approach. It attempts to optimize the student's mastery of domain knowledge by generating new problems, concepts, and instruction feedback. e-Tutoring systems are computer-based teaching systems that regularly give students quick and personalized guidance or feedback without mediation from a personal teacher. e-Tutoring is an excellent subject in the context of e-learning. e-Learning environments are increasingly getting popular in different contexts in academies, universities, and vocational training [4]. Therefore, suitable support of learners is also getting great significance. Besides, collaborative learning is also growing, which puts greater demands on learners, especially when the collaboration is implemented and tailored to enhance learning and learning results [4].

The area of instruction is one of the first areas in which ontologies are applied in as a cognitive tool. For several considerations, this was a consequence to the extensive adoption of the constructivist paradigm of education and the general application of such knowledge technologies as concept maps, mind maps, and several more for learning goals [5]. Recently, scientists found that the benefits of using ontologies in the educational field have relative support for designing the coming e-learning environment. The term "ontology" comes from philosophy and it is defined as "a set of representational primitives with which to model a domain of knowledge or discourse" [6]. So, an ontology gives a specific view on some part of the world. While knowledge representation formalisms define how to represent concepts, ontologies identify the concepts to be described and are linked. Thus, ontology can be noticed as a well-established and widely agreed-upon system of concepts in a specific knowledge domain and their relationship. Specialized knowledge domain ontologies can be utilized as a semantic pillar for topics or repositories of teaching resources. By offering agreed-upon terminologies for the domain knowledge representation, ontologies would help

share, reuse, and exchange topic units. Ontologies also enable the availability of machine-readable web resources.

Most of the current solutions of e-tutor systems are designed and developed for a particular domain which means the provided solution will not be suitable for another knowledge domain. Therefore, these systems developed for isolated knowledge bases have some limitations and drawbacks for using local knowledge bases. These limitations refer to lack of standardization, limited knowledge base shareability, lack of flexibility, lack of reusability, and manual control. Due to these shortcomings, an ontology domain knowledge model is proposed to avoid the limitations and drawbacks of using local knowledge bases. To improve and increase the learning quality and process, a novel ontology domain knowledge model develops by defining a set of relationships that would be adequate and clear to represent all possible relationships for developing and building the ontology domain knowledge model. In the proposed ontology domain knowledge model, two types of domain ontologies were introduced: a) general concepts for domain knowledge model ontology and b) specific domain knowledge model ontology. The general concepts of the proposed ontology domain knowledge model deal with the domain knowledge model concepts and define the relationship related to these concepts. The ontology of a selected domain knowledge model deals with the selected subject area that can relate the selected subject domain to the general concepts for the domain knowledge model. It seems like individuals or instances for the general concepts of the domain knowledge model.

Another key issue in current systems is the limited concept level functionality. Many systems provide data-level features as they lack the information and engine to perform smart operations. The proposed ontology framework includes reasoning engines that can improve the adaptivity, customization of the e-tutor frameworks. In the proposed knowledge base, the schema can contain also rules as part of the background ontology. This approach supports a declarative description model instead of the procedural way; thus, it has an increased flexibility and coding efficiency.

This paper constructs a domain knowledge model for an e-tutor system which in turn would help enhancing teaching and learning process. Furthermore, the implementation of this model is described in Python, which can be applied in the future to support the problem-solving process.

This article is organized as follows. The introduction and methodology of designing the ontology are explained in the first and second sections. The third section displays the related work, the fourth section illustrates the proposed ontology of the domain knowledge model and explains the proposal model in detail, the fifth section presents a case study on ontology implementation using Python and Owlready2 module, the sixth section demonstrates the result and discussion, and the conclusion is in the seventh section.

II. METHODOLOGY

The primary goal of this work is to enhance the quality of the learning process by making it in a personalized way. The process design and development of ontology usually

encompasses several standard tasks. However, there is no dominating approach for constructing the ontologies. The main principle is to define the ontology concepts which are related to the objects and the relationships for the selected domain. The methodology for creating and building an ontology assumes defining the objectives and domain of applicability. Moreover, it must be identified in higher-level details: what is the purpose of designing the domain ontology, what are the types of questions that should be answered through it, how it will be utilized and supported for solving the problem for the selected domain. Several techniques for the design and development of ontology are given such as [7] and [8]. Though these methods are somewhat different and are influenced in varying ways by the technology used, the underlying processes of developing the domain ontology are similar. Therefore, the suggested ontology development process is composed of the following phases:

- Domain and purpose of the ontology.
- Discover if there are related ontologies.
- Enumerate important terms in the domain.
- Defining the key classes and their hierarchy.
- Identify the properties of classes.
- Facets attaching to properties.
- Creating class instances.

III. RELATED WORK

The technique known as e-tutoring system has been tracked by scholars in education, psychology, and artificial intelligence. The aim of e-tutoring system is to offer the advantages of one-to-one teaching. It allows students to train their skills by bringing out activities in greatly interactive learning platforms. e-Tutoring system is a computer application that uses artificial intelligence approaches to improve and personalize the teaching and learning processes [9]. e-Tutoring differs from other e-learning systems because it uses a knowledge base to guide the pedagogical approach. It attempts to optimize the student's mastery of domain knowledge by generating new problems, concepts, and instruction feedback. There are many e-tutoring systems meant and developed for learning and teaching reasons. These systems support learners to progress quickly and improve their self-confidence. SmartTutor is a web-based intelligent tutoring system developed to support teaching and students based on knowledge background, skills, and teaching techniques [10]. The remarkable aspect of SmartTutor is the incorporating of instructional and artificial intelligence (AI) methods in a unified intelligent e-tutoring system to give personalized help to learners based on the knowledge-level of individual learners [11]. e-Tutoring systems are developed to illustrate the essential knowledge on the subject, inform which kinds of knowledge can be used to solve problems in the given domain, generate and offer suitable task activity based on the recorded performance, and recommend the next suitable task for the learner to choose. Smart Tutor delivers the capability of a tutor to adjust to individual learners' requirements and skills [10]. e-Tutoring systems, also known as Intelligent Tutoring Systems

(ITS), became a reality in the mid-1970s and peaked at the end of the 1980s, when expert systems were in use [12]. During the nineties, most artificial intelligence and education investigations concentrated on intelligent learning environments more influenced by Computer-based learning, microworlds, and Computer-based training [13]. Of course, ITS study is still an active sub-field of research on adaptive learning. ITS systems were not cost-effective enough to survive in education and training. ITS got interested in the late 2010s, particularly since "statistical AI" based on neural networks and relevant methods allows extracting patterns from big data. These can be due to recommender systems that recommend learning analytic systems that detect learners' difficulties. ITS are adaptive systems that use intelligent technologies to tailor learning based on various characteristics of students, including their background knowledge, mood, and learning style. The e-tutoring system includes three kinds of knowledge to give students proper education, arranged into four separate system modules (as shown in Fig. 1).

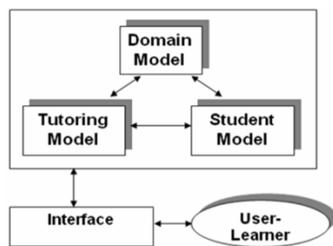


Fig. 1. Traditional Components of an ITS [10].

The tutor model is a computer model that helps the system choose the appropriate explanation style for a particular learner based on the learner's history tracked by the learner model. This knowledge allows the e-tutoring system to check the student's responses and steps with an expert to assess what they know and do not know. The student model is the learner's knowledge-level while the student works on e-tutoring system. The model assesses student's behavior while working with the e-tutoring system to find their knowledge, cognitive skills, and reasoning abilities. The domain model is concerned with the knowledge associated with the problem considered for a specific domain, including the teaching materials and the meta-information about the topic to be taught. It depends on the learner model's diagnostic processes to determine what, when, and how to display information to a learner. The interface model is essential as a delivery mechanism and learning facility to help the learner complete the activity. It can also perform as an external representation of the expert and teaching models.

Researchers have investigated practices of knowledge representation such as semantic-based, rule-based, case-based, frame-based, Bayesian network, logic-based, and ontology-based. Rule-based models are also called Cognitive tutors. The rule-based models are built from cognitive task analysis, producing problem spaces or task models. These problem spaces or task models are constructed by observing the expert and novice users. Task models represent a set of production rules in which each rule represents an action corresponding to a task [14]. When a user tries to solve a given task, the user's reasoning ability is analyzed based on the rules applied by the

user, i.e., the user's solution is compared step-by-step to the solution given by the expert. Case-based is an artificial intelligence problem-solving method that records experience into cases and associates the current problem with an experience [15]. A logical representation language has some definite rules for dealing with propositions and reasoning for knowledge and representation. Logical representation entails deducing a conclusion from many circumstances. This representation establishes many fundamental communication principles. It is composed of well-defined syntax and semantics that facilitate sound inference. Each phrase can transform into a logical form through syntax and semantics. Frame-Based is one of the artificial intelligence techniques to structure the data, and it is used to separate information into substructures via the representation of stereotyped scenarios [16]. Frame-Based seems like a record form build-up of many characteristics and values used to describe an object in the real world. In addition, this object contains a collection of slots and their associated values. However, these slots come in a combination of shapes and sizes. Facets are the names and values assigned to slots. A Bayesian network is also referred to as a belief network or Bayes net. Bayesian network is probabilistic and graphical in a form of graph with directed acyclic devoid of loops and self-connections used for knowledge representation for an uncertain domain, with each node indicating a random variable [17]. Each edge denotes a conditional probability associated with the associated random variables. Semantic-based is a knowledge representation method that enables visualization of the knowledge via graphical networks [18]. This network incorporates nodes representing entities and arcs that reflect their relationships. Moreover, Semantic-based classify things in different ways and can also connect them in the style of a graph.

Ontologies, as semantic-based representation, have gained vital significance as one of the most commonly used techniques to describe and share knowledge in several disciplines such as E-learning systems, business modeling, software engineering, knowledge engineering [19].

Regarding the continuous development of new technology, it can change the way of teaching and learning. According to Fensel [20], the primary reason for the popularity of ontologies is due to providing "a shared and common understanding of a domain that can be communicated between people and application systems. Ontology can be constructed as a representation required for scale and variety in the design of educational frameworks. In the e-learning field, ontologies are employed in various applications extending from domain knowledge modules representation to automate generation and assessment of personalized learning materials. The concept of ontology is a useful technology that incorporates related resources, shares knowledge, and eliminates unnecessary data. Ontology is a fundamental description of the information in the world [21]. The ontology in computing refers to knowledge representation applying a collection of concepts and connections among them [22]. In the context of a targeted discipline, ontology is used to rationally reason and validate concepts in the semantic knowledge model. In theory, ontology is a "formal, explicit specification of a shared conceptualization" [23]. It offers a shared vocabulary that can

be employed to construct the domain knowledge model, involving objects, concepts, properties, and relationships.

A comparison was given utilizing some selected criteria according to the knowledge representation models employed in the present works for representing the domain knowledge module as a part of an E-tutor framework. The authors considered some criteria for comparing the representation model used with others in the literature. The suggested criteria covered a number of terminologies: standardability, reusability, flexibility, shareability, simplicity, and reasoning engine, as indicated in Table I.

TABLE I. COMPARISON OF THE MODELS

Models	Criteria					
	Standard ability	Reusability	Flexibility	Shareability	Simplicity	Reasoning engine
Semantic-based	√	√	×	√	×	√
Rule-based	√	√	×	×	×	√
Case-based	√	×	×	×	×	√
Frame-based	√	√	√	×	×	√
Bayesian network	√	×	×	×	×	√
Logic-based	√	×	×	√	×	√
Ontology-based	√	√	√	√	√	√

√ means feature is allowed, and × means feature not allowed.

IV. PROPOSED DOMAIN KNOWLEDGE MODEL FOR E-TUTOR FRAMEWORK

The domain model is the system's knowledge base, and it organizes the domain knowledge structure, its various key concepts, and the relationships between the concepts. This model essentially deals with the what-to-teach part of e-tutoring system [24]. The domain model is concerned with domain knowledge construction, organization, topics, and relationships [24]. Domain knowledge is a set of suggestions that identify all the vocabulary concepts to explain or solve problems. Domain knowledge is only declarative, and it does not tell how learners can use the domain knowledge model to solve a practical problem [25].

Based on the properties of the learning materials, two kinds of ontologies are employed, and these are general concepts domain knowledge ontology and specific domain knowledge ontology. These modules represent the knowledge to be learned, deliver input to the expert model, and eventually provide specific feedback, select problems, generate guidance, and support the student model.

A novel domain knowledge model was suggested based on the current research area, as shown in Fig. 2. This model is based on topics, concepts, attributes, tasks, competencies, assessment, and relations. In order to share and reuse the domain knowledge model in e-tutoring systems, ontologies are employed to organize and represent the domain knowledge

model. The benefit of this model is to personalize the materials for learners.

Based on the general concepts of the domain knowledge ontology shown in Fig. 2, topics, concepts, attributes, tasks, competencies, and assessment terms refer to the following:

- Topics can be utilized to present domain knowledge or a comprehensive overview of a subject or course.
- The concept identifies the sub-domain or unit of a subject or course.
- Competency is used to demonstrate the features and skills that allow and enhance the efficiency of student performance to gain new knowledge and understand specific topics.
- The task is used to demonstrate how a student can complete a task within a given period of time.
- The attribute represents a topic or domain attribute within a domain model.
- The assessment is used to present how the system can evaluate or assess the student activities required within a given period of time.

Fig. 3 displays the design of a specific domain knowledge ontology case study for IT domain in e-tutor system for the computer programming. Many types of relationships are used in the selected case study, such as specialization or generalization, association, and containment. A containment means that a specific topic within a domain contains different concepts (has-a). The specialization or generalization means that certain topics or domains have specific concepts (is-a). The association means that a specific topic or concepts associate with each other. Based on Fig. 2 and Fig. 3, the following shows a brief description of a subject:

- Topic: Control Structure.
- Concept: Loop, Sequence and Condition.
- Competency: understand, analyze, implement.
- Task: program, code review, project.
- Attribute: syntax, operators.
- Assessment: activities such as quizzes, tests.

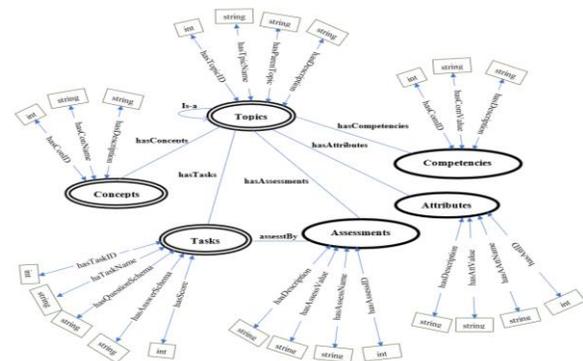


Fig. 2. The Proposed Ontology-based Model for Domain Knowledge.

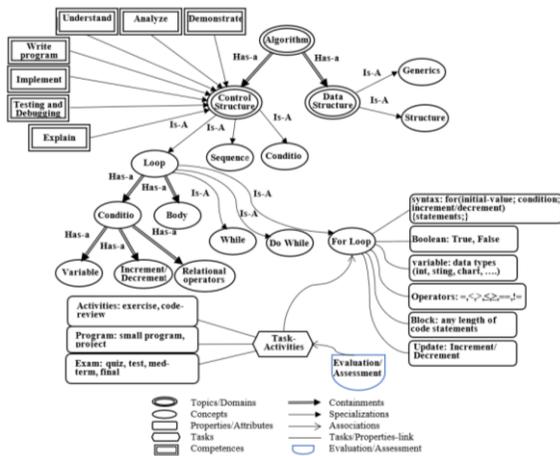


Fig. 3. Domain Knowledge Module Instances for e-tutoring Frameworks.

V. IMPLEMENTATION OF THE PROPOSED MODEL

Information science and technology provides many modules and packages for ontology construction and management. Python is one of the most popular languages adopted when implementing an ontology for the domain knowledge model. It is an interpreted, object-oriented, extensible programming language [23], which provides an excellent combination of clarity and versatility in different disciplines. The domain knowledge model, considered here, is “Basics of Computer Programming”, and the ontology created consisted of the “Algorithm in Computer Programming”. Fig. 4, 5, 6, 7, 8, 9, and 10 show the implementation of the proposed model using Python.

Fig. 4 displays a snippet of the domain knowledge components construction using Python and Owlready2 syntax "with ontology: ..." to indicate that the constructed ontology will receive the new RDF triples and the class keyword for declaring the components of the ontology. While Fig. 7 demonstrates the output of the ontology of the domain knowledge model components as a list of concepts. Fig. 5 displays a snippet of the object property related to the domain knowledge model in the format of domain and range. For example, "hasParts(Topics >> Concepts)" hasParts is the object property relationship that means each Topic may contain some Concept in this example, Topics is the domain of the relationship, and Concepts is the range of the relationship. Fig. 6 displays a snippet of the data property relationship related to the domain knowledge model in the structure of domain and range. Considering "topicName(Topics >> str, FunctionalProperty)," topicName declares the name of the Topics component, Topics are the domain, and str (means the topicName datatype is a string) is the range of the relationship. Fig. 8 presents how to insert the "While Loop" topic data as an instance related to the Topics component. Checking the consistency of the ontology using the Hermit reasoner, the most commonly used in ontology engineering applying sync_reasoner(), is shown in Fig. 9. Rule construction in SWRL layouts for inferring further knowledge can add to the created domain knowledge model of the selected ontology components using imp = Imp() and imp.set_as_rule() functions, as shown in Fig. 10.

```
# Domain Knowledge Model Components
with ontology:
class Topics(Thing):
def take(self):
print('I took a topic domain')
class Concepts(Thing):
def take(self):
print('Concepts related to the topic domain')
class Tasks(Thing):
def take(self):
print('Tasks related to the topic domain')
class Competencies(Thing): pass
class Attributes(Thing):
def take(self):
print('Competencies related to the topic domain')
class Assessments(Thing):
def take(self):
print('Assessments of the tasks related to the topic domain')
```

Fig. 4. Domain Knowledge Components Construction.

```
#Object Property of the Domain Knowledge Model
class hasParts(Topics >> Concepts): pass
class partsOf(Concepts >> Topics):
inverse = hasParts
class hasCompetencies(Topics >> Competencies): pass
class competenciesOf(Competencies >> Topics):
inverse = hasCompetencies
class hasParents(Topics >> Topics): pass
class parentsOf(Topics >> Topics):
inverse = hasParents
class hasAttributes(Topics >> Attributes): pass
class attributesOf(Attributes >> Topics):
inverse = hasAttributes
class hasTasks(Topics >> Tasks): pass
class tasksOf(Tasks >> Topics):
inverse = hasTasks
```

Fig. 5. Object Property of the Domain Knowledge.

```
48 #Topic Data Property
49 class topicID(Topics >> int, FunctionalProperty):
50     pass
51 class topicName(Topics >> str, FunctionalProperty):
52     pass
53 class parentTopic(Topics >> str, FunctionalProperty):
54     pass
55 class topicDescription(Topics >> str, FunctionalProperty):
56     pass
```

Fig. 6. Data Property of the Domain Knowledge.

```
Out[3]: [DKMontology.Topics,
DKMontology.Concepts,
DKMontology.Tasks,
DKMontology.Competencies,
DKMontology.Attributes,
DKMontology.Assessments]
```

Fig. 7. The List of Concepts.

```
topic11 = Topics(
'topic11',
topicID = 11,
topicName = 'while Loop',
parentTopic = 'loop',
topicDescription = '''In most computer programming languages, a while loop is a control flow statement that allows
code to be executed repeatedly based on a given Boolean condition. The while loop can be thought of as a repeating
if statement. The while loop can be thought of as a repeating if statement.'''
)
```

Fig. 8. A Topic Instance for While Loop.

```
3 try:
4     sync_reasoner()
5     print("\nOk, the constructed ontology is consistent and allows the classification, instance checking,
6     class satisfiability, and conjunctive query answering.")
7 except OwlreadyInconsistentOntologyError:
8     print("\nthe constructed ontology is inconsistent and didn't allow the classification, instance checking,
9     class satisfiability, and conjunctive query answering.")

* Owlready2 * Running Hermit...
java -xms2000M -cp C:\Users\Muhammad\anaconda3\lib\site-packages\owlready2\hermit;c:\Users\Muhammad\anaconda3\lib\site-packages\owlready2\hermit\hermit.jar org.semanticweb.Hermit.cli.CommandLine -c -D -I file:///C:/Users/Muhammad/anaconda3/ta/local/temp/tmp0v05se

Ok, the constructed ontology is consistent and allows the classification, instance checking,
class satisfiability, and conjunctive query answering.

* Owlready2 * Hermit took 3.44258967025757 seconds
* Owlready * Reparenting history.world_war_II: (history.Topics) => (history.TaskAssessments, history.Topics, history.Attribute)
* Owlready * Reparenting history.world_war_I: (history.Topics) => (history.TaskAssessments, history.Topics, history.Attributes)
* Owlready * Reparenting history.civil_war: (history.Topics) => (history.TaskAssessments, history.Topics, history.Attributes)
* Owlready * (NB: only changes on entities loaded in python are shown, other changes are done but not listed)
```

Fig. 9. Checking the Consistency of the Ontology.

```

1 with ontology:
2   imp = Imp()
3   imp.set_as_rule("""
4     Topics(?t, hasTasks(?t, ?s), assesstBy(?t, ?s), hasCompetencies(?t, ?co) ->
5     hasAssessments(?t, ?s)
6     """)

1 sync_reasoner_pellet(infer_property_values = True, infer_data_property_values = True)

* Owlready2 * Adding relation DKM.assessment2.assessmentValue small program
* Owlready2 * Adding relation DKM.assessment2.assessmentDescription ()
* Owlready2 * Adding relation DKM.task3.answSchema Loops make it unnecessary to repeat a process in an algorithm. Instead
* Owlready2 * Adding relation DKM.attribute4.attributeValue =

* Owlready2 * Pellet took 2.9846913814544678 seconds
* Owlready2 * (NB: only changes on entities loaded in Python are shown, other changes are done but not listed)

```

Fig. 10. Rule Construction and the Results of Adding New Knowledge.

Protégé is used to display the ontology graph of while loop topic shown in Fig. 11. A SPARQL query about displaying all the Topics in the developed ontology domain knowledge model, considering retrieving the topic “While” and its description is demonstrated in Fig. 12.

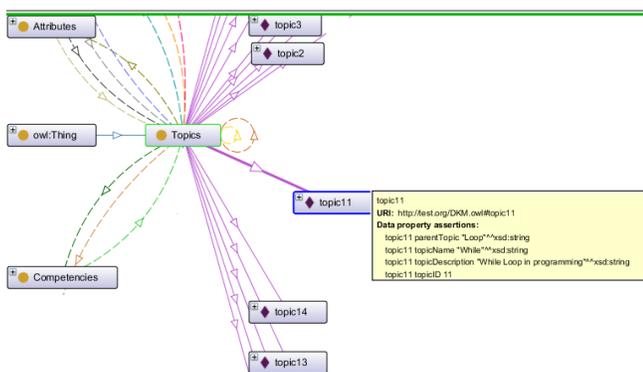


Fig. 11. The Ontology Graph of While Loop Topic.

```

result = """ PREFIX topic: <http://test.org/OntologyDKM.owl#>
SELECT ?topicID ?topic ?Description
WHERE {
  ?t a topic:Topics;
  topic:topicName ?topic;
  topic:topicDescription ?Description .
}
ORDER BY ?topicID DESC(?topicID)"""

query_result = g.query(result)
for query in query_result:
    print('')
    print(f"Topic Name: {query.topic} \nTopic Description: {query.Description}")

Topic Name: While
Topic Description: In most computer programming languages, a while loop is a control flow statement that allows code to be executed repeatedly based on a given Boolean condition. The while loop can be thought of as a repeating if statement. The while loop can be thought of as a repeating if statement.

```

Fig. 12. A SPARQL Query to Retrieve the Topic “While” and its Description.

VI. RESULT AND DISCUSSION

The most commonly utilized knowledge representations are semantic-based, rule-based, case-based, frame-based, Bayesian network, logic-based, and ontology-based [14], [15], [16], [17], [18] approaches. A comparison of knowledge representation forms is conducted according to some criteria or features covered: standardability, reusability, flexibility, shareability, simplicity, and reasoning engine. Ontologies, as knowledge-based representation, have gained vital significance as one of the most commonly used techniques to describe and share knowledge in several disciplines such as E-learning systems, business modeling, software engineering, knowledge engineering [19].

The semantic web is an extension of the World Wide Web, is crucial in the development of personalized learning in e-learning systems. Semantic web technologies such as RDF, XML, and ontologies can be used for knowledge representation and reasoning of the teaching materials. Ontologies, which can

be defined as a specification of a conceptualization, have the advantage of addressing the challenges of interoperability between educational repositories of various e-learning systems.

An ontology-based domain knowledge model has been proposed. In section 4, a theoretical model is described based on two kinds of ontologies. First, a general concepts domain knowledge ontology based on topics, concepts, attributes, tasks, competencies, assessment, and relations is presented in Fig. 2. Second, a specific domain knowledge ontology is designed as a case study for IT domain in programming using different types of relationships such as specialization or generalization, association, and containment as depicted in Fig. 3. In Section 4, an implementation of the ontology-based model is delivered using Python, Owlready2, and Protégé.

The current work deals with declarative knowledge that helps students understand all the vocabulary concepts in a specific subject. The future work suggests dealing with problem-solving process that works on procedural knowledge to help students in understanding how to use domain knowledge in solving a practical problem.

The proposed model is introduced to explain how the ontology domain knowledge model can be combined with an e-tutor system to improve the quality of intelligent problem-solving. Also, it will make it possible to reuse knowledge components and develop e-tutor system frameworks. Finally, a domain knowledge model for e-tutoring system, which can serve to enhance teaching and learning, is proposed. Furthermore, this model can avoid the issue currently exists in intelligent e-learning systems related to isolating knowledge bases. The majority of the domain knowledge models use an isolated knowledge base, and this local knowledge base can only provide limited knowledge background. The limitations of the isolated knowledge base are lack of standardability, reusability, flexibility, and limited knowledge. The solution can satisfy the characteristics of reusability, standardability, open knowledge, and flexibility.

By using ontology as a knowledge representation technique for building the domain knowledge model, the problem of isolated knowledge bases can be avoided. The developed ontology can be involved to manage adaptive intelligent e-learning frameworks in the future.

VII. CONCLUSION

An intelligent tutoring system or E-tutoring System is a type of educational system that uses artificial intelligence to reflect knowledge. The use of e-tutoring systems has become a key component of enhancing educational activities. In this paper, readers can see that e-tutoring is very significant in the context of the growing number of e-learning frameworks. An ontology-based knowledge model for IT domain has been developed for e-tutor system.

This paper constructs a domain knowledge model for an e-tutor system which in turn would help enhancing teaching and learning process. Furthermore, the implementation of this model is described in Python and Owlready2 module, which can be applied in the future to support the problem-solving process. Two types of ontologies are employed, and these are general concepts domain knowledge ontology and specific

domain knowledge ontology. These modules represent the knowledge to be learned, deliver input to the expert model, and eventually provide specific feedback, select problems, generate guidance, and support the student model.

The current work deals with declarative knowledge that helps students understand all the vocabulary concepts in a specific subject. The future work suggests dealing with problem-solving process that works on procedural knowledge to help students in understanding how to use domain knowledge in solving a practical problem.

ACKNOWLEDGMENT

The described study was carried out as part of the EFOP-3.6.1-16-00011 “Younger and Renewing University – Innovative Knowledge City – institutional development of the University of Miskolc aiming at intelligent specialization” project implemented in the framework of the Szechenyi 2020 program. The realization of this project is supported by the European Union, co-financed by the European Social Fund.

REFERENCES

- [1] R. Prakash and S. Kumar, “E-LEARNING FRAMEWORK FOR SEMANTIC-WEB,” vol. 12, no. 14, pp. 2994–3006, 2021, [Online]. Available: <https://turcomat.org/index.php/turkbilmat/issue/view/49>.
- [2] A. Kumar and N. J. Ahuja, “An adaptive framework of learner model using learner characteristics for intelligent tutoring systems,” in *Advances in Intelligent Systems and Computing*, 2020, vol. 989, pp. 425–433, doi: 10.1007/978-981-13-8618-3_45.
- [3] “Artificial Intelligence in Education (AIED) a high-level academic.pdf,” *AI Ethics*, Springer, vol. 2022, no. 2, pp. 157–165, 2021, doi: <https://doi.org/10.1007/s43681-021-00074-z>.
- [4] F. Ç. Baz, “New Trends in e-Learning,” in *Trends in E-learning*, M. Sinecen, Ed. Rijeka: IntechOpen, 2018.
- [5] D. Tomar and P. Tomar, “Artificial Intelligence-Based Knowledge Representation and Reasoning,” in *Impact of AI Technologies on Teaching, Learning, and Research in Higher Education*, S. Verma and P. Tomar, Eds. IGI Global, 2021, pp. 134–149.
- [6] T. Gruber, “Ontology,” in *Encyclopedia of Database Systems*, L. LIU and M. T. ÖZSU, Eds. Boston, MA: Springer US, 2009, pp. 1963–1965.
- [7] E. Katis, H. Kondylakis, G. Agathangelos, and K. Vassilakis, *Developing an ontology for curriculum and syllabus*, vol. 11155 LNCS, no. 1. Springer International Publishing, 2018.
- [8] H. El-Ghalayini, “E-course ontology for developing E-learning courses,” in *2011 Developments in E-systems Engineering*, 2011, pp. 245–249.
- [9] P. Sharma and M. Harkishan, “Designing an intelligent tutoring system for computer programing in the Pacific,” *Educ. Inf. Technol.*, no. 0123456789, 2022, doi: 10.1007/s10639-021-10882-9.
- [10] A. Khan, M. Junaid, and M. Noman, “Smart Tutor an Intelligent Tutoring System for C Sharp Programming Bahria University Karachi Campus,” *Int. J. Comput. Appl.*, vol. 180, no. 27, pp. 28–33, 2018, doi: 10.5120/ijca2018916646.
- [11] R. Pavlov, “Smart Tutors : Delivering Personalized Learning,” pp. 1–4, 2021, [Online]. Available: <https://www.epam.com/insights/blogs/smart-tutors-delivering-personalized-learning>.
- [12] T. Crow, A. Luxton-Reilly, and B. Wuensche, “Intelligent Tutoring Systems for Programming Education: A Systematic Review,” *ACM Int. Conf. Proceeding Ser.*, pp. 53–62, 2018, doi: 10.1145/3160489.3160492.
- [13] L. Chen, P. Chen, and Z. Lin, “Artificial Intelligence in Education: A Review,” *IEEE Access*, vol. 8, pp. 75264–75278, 2020, doi: DOI:10.1109/ACCESS.2020.2988510.
- [14] N. Mendjoge;, A. R. Joshi;, and M. Narvekar, “Review of knowledge representation techniques for Intelligent Tutoring System,” in *2016 International Conference on Computing for Sustainable Global Development (INDIACom)*, 2016, pp. 2508–2512.
- [15] M. Masood and N. A. M. Mokmin, “Case-based reasoning intelligent tutoring system: An application of big data and IoT,” *ACM Int. Conf. Proceeding Ser.*, vol. Part F132530, pp. 28–32, 2017, doi: 10.1145/3152723.3152735.
- [16] W. Abu-Dawwas and M. Abu-Dawas, “Proposed frame-based expert system to construct student’s knowledge model in intelligent tutoring systems,” *J. Math. Comput. Sci.*, vol. 10, no. 5, pp. 1529–1537, 2020, doi: 10.28919/jmcs/4567.
- [17] A. Tato, R. Nkambou, J. Brisson, C. Kenfack, S. Robert, and P. Kissok, “A Bayesian network for the cognitive diagnosis of deductive reasoning,” *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 9891 LNCS, pp. 627–631, 2016, doi: 10.1007/978-3-319-45153-4_78.
- [18] M. Khfagy, Y. Abdelsatar, and O. Reyad, “Knowledge Representation in Intelligent Tutoring System,” vol. 533, no. 1, 2017, doi: 10.1007/978-3-319-48308-5.
- [19] J. Alshboul, H. A. A. Ghanim, and E. Baksa-Varga, “Semantic Modeling For Learning Materials In E-Tutor Systems,” *J. Softw. Eng. Intell. Syst.*, vol. 6, no. 2, pp. 17–24, 2021, [Online]. Available: <https://www.jseis.org/volume/vol6no2.html>.
- [20] A. Verma and A. Verma, “An Ontology of Ontological Engineering,” *Int. J. Eng. Sci.*, vol. 24, no. 63019, pp. 97–107, 2017.
- [21] C. Ma and B. Molnár, “Use of Ontology Learning in Information System Integration: A Literature Survey,” *Communications in Computer and Information Science*, vol. 1178 CCIS, pp. 342–353, 2020, doi: 10.1007/978-981-15-3380-8_30.
- [22] S. Chimalakonda and K. V. Nori, “An ontology based modeling framework for design of educational technologies,” *Smart Learn. Environ.*, vol. 7, no. 1, 2020, doi: 10.1186/s40561-020-00135-6.
- [23] T. Guber, “A translational approach to portable ontologies,” *Knowl. Acquis.*, vol. 5, no. 2, pp. 199–229, 1993.
- [24] A. Ramírez-Noriega et al., “Towards the Automatic Construction of an Intelligent Tutoring System: Domain Module,” *Adv. Intell. Syst. Comput.*, vol. 930, no. 3, pp. 293–302, 2019, doi: 10.1007/978-3-030-16181-1_28.
- [25] W. Yathongchai, J. Angskun, and C. C. Fung, “An Ontology Model for Developing a SQL Personalized Intelligent Tutoring System,” *Naresuan Univ. J. Sci. Technol.*, vol. 25, no. 4, pp. 88–96, 2017.

The Design of Home Fire Monitoring System based on NB-IoT

Jun Wang, Ting Ke, Mengjie Hou, Gangyu Hu
School of Mechanical and
Electrical Engineering
Jiangxi University of
Science and Technology
Ganzhou, China

Abstract—In the field of home fire monitoring, the currently relatively mature monitoring solutions include GPRS/GSM communication and Zigbee communication. The main disadvantage of GPRS wireless communication is high power consumption, and the disadvantage of Zigbee technology is that it needs to be combined with other communication technologies to realize remote monitoring. In addition, the above technical solutions all require self-built local or remote monitoring servers to save monitoring data. In view of the above problems, this system designs a home fire monitoring system based on NB-IoT technology and cloud platform. The system uses a single-chip STM32F103C8T6 as the core controller and contains a sensor data acquisition module and a narrowband IoT communication module. The data fusion of multi-sensor data is performed by BP neural network algorithm. On the basis of remote transmission, the system solves the problems of high power consumption, high cost and insufficient signal coverage of terminal hardware. The system can collect indoor environmental parameters and fire information in real time, and upload them to the cloud platform for storage. If abnormal data is detected, an early warning message will be issued. The feasibility of the system is verified, and the verification results show that the system works normally and the output is accurate, which meets the design requirements and can be widely used.

Keywords—NB-IoT; cloud platform; fire monitoring system; STM32F103C8T6; sensor; BP neural network

I. INTRODUCTION

Nowadays, high-tech electronic products are widely used in various families. However, because some people can't use these high-tech products reasonably, and even misoperation leads to adverse consequences, many abnormal situations or dangers occur. In addition, most families use natural gas, liquefied petroleum gas, etc. Some people forget to close the valve after use, causing gas leakage and other situations. Gas is a flammable and explosive gas, which is likely to cause fire or explosion accidents. Therefore, the indoor system needs to have the functions of gas detection and fire detection to ensure the safety of family environment. The establishment of family fire monitoring system has practical application value and social benefits [1].

In order to solve the above safety problems, a home fire monitoring system based on NB-IoT (Narrow Band Internet of Things) is designed. By using a variety of heterogeneous sensors (such as temperature and humidity sensors, smoke sensors, etc.) as the data acquisition module to collect indoor environmental information, the data is analyzed and processed

by the main controller to realize the functions of fire detection and gas leakage detection, so as to achieve the purpose of remote care of the home environment, ensure the safety of family life and property, and provide a safe and intelligent living environment for the family [2].

The core controller of the home fire monitoring system is a microprocessor STM32F103, which contains a sensor data acquisition module and a Narrowband IoT Communication Module composed of a home living room environment monitoring and fire monitoring system [3]. The system is based on the middle layer of the home fire monitoring system of Huawei's OceanConnect platform, and achieves the work of device data reporting and platform issuing commands [4]. The home fire monitoring platform realizes the remote monitoring of the home living environment. The validity of the fire detection algorithm is verified through experiments, and the test results are as expected; through the test of the terminal node, cloud platform and each functional module in the home fire monitoring platform, the feasibility of the function of the home fire monitoring system and the stability of data transmission are verified, which provides a reference for the practical application and promotion of the system.

II. RELATED WORK

FENG Hui et. al. [5] based on ZigBee technology, smoke, CO, temperature detection fusion were integrated, and community fire alarm system based on ZigBee is designed and implemented. QI Bin et. al. [6] used LoRa and GPRS separately for long- distance transmission of fire sensing information and fire alarm information. The combination of the two technologies meets the needs of wireless fire alarm system monitoring and alarming. ZHANG Zhi-hua et. al. [7] used probabilistic neural networks for information fusion of fire features, which can effectively perform fire identification and improve the accuracy of fire detection. XIE Rongquan et. al. [8] used the BP neural network algorithm to calculate and detect the developing rule and signal feature of the fire image. OKOKPUJIE K O et. al. [9] combined GPRS and single-chip microcomputer to build an automatic fire alarm system to realize remote fire monitoring.

Although Zigbee technology has flexible terminal nodes, easy deployment, and low power consumption, its application distance is limited. The disadvantages of GPRS/GSM wireless communication method are high power consumption, high operation and maintenance costs, low data transmission rate,

and the risk of withdrawing from the network with the rapid development of 5G technology in China. LoRa wireless communication technology needs to be combined with GPRS/GSM wireless communication method to achieve remote monitoring. In addition, the above technical solutions all require self-built local or remote monitoring servers to save monitoring data.

III. OVERALL SYSTEM ARCHITECTURE DESIGN

The system consists of sensing terminal, cloud platform and application layer. The sensing terminal mainly contains sensors, STM32 main control module, NB-IoT communication module and external control devices; the platform layer mainly contains IoT cloud platform; the application layer provides home fire monitoring application services to users, and the system architecture is shown in Fig. 1.

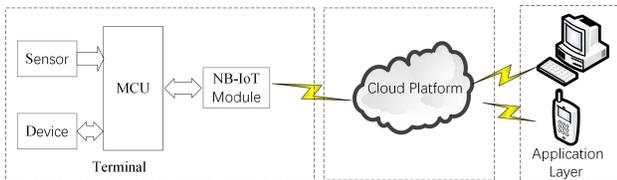


Fig. 1. Overall System Structure Diagram.

The sensing terminal collects indoor temperature and humidity, smoke concentration, harmful gases and other information, uploads alarm signals when abnormal conditions such as gas leakage and fire occur, sends the data to the cloud platform, and starts external control devices at the same time.

The IoT cloud platform is essentially a cloud server with functions of device management, data management, etc. It converts the format of upstream and downstream service data, connects NB-IoT terminal devices in the south direction, and connects the application layer in the north direction, decodes the south direction data for easy access by the perception layer subscription, and encodes the north direction commands for easy reception by the perception layer.

The northbound application is a human-computer interaction interface, which accesses data messages through a personal PC subscribed by the API interface and protocols opened by the IoT cloud platform, displays the home environment data status in real time, and includes alarm notification functions, and can issue commands to remotely control terminal devices when abnormal.

IV. TERMINAL HARDWARE DESIGN

The sensing terminal mainly consists of a variety of heterogeneous sensors, the main controller, NB-IoT module and other peripheral circuits, in which the sensors are mainly responsible for collecting indoor environmental data, the main control unit fuses, processes and controls the start and stop of the collected data, and establishes a connection with the cloud platform through the NB-IoT module networking, packages and uploads the data to the cloud platform, and receives commands from the cloud platform. The hardware block diagram of the system terminal is shown in Fig. 2.

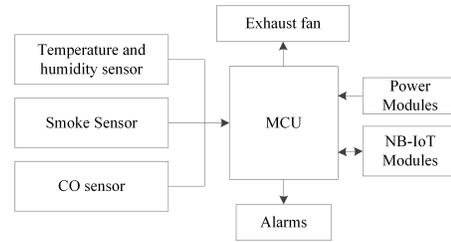


Fig. 2. Terminal Hardware Block Diagram.

A. Environmental Parameter Acquisition Circuit

The system uses STM32F103C8T6 as the terminal main control chip, which has timer, UART, ADC, I/O and other modules inside, which fully meets the functional requirements of system hardware design. The temperature and humidity acquisition module DHT11 is used to monitor changes in temperature and humidity, the smoke sensor MQ-2 is used to detect gas leakage, and the carbon monoxide sensor 4CO-500 is used to detect CO concentration [10].

- 1) DHT11 chip adopts single bus data format. With high measurement accuracy and low power consumption, it only needs a single data pin port to complete I/O bidirectional transmission. Connect the I/O port PA7 of the STM32 microcontroller to the DATA pin of DHT11, and connect a pull-up resistor to the DATA pin to realize temperature and humidity acquisition.
- 2) The system uses MQ-2 smoke gas sensor to monitor the environment for gas leaks and fires, which is commonly used to detect smoke, liquefied gas, alcohol, methane and other gases, with the advantages of high sensitivity, wide detection range, high stability, long service life, etc. and is widely used in smart homes and other fields. The signal output by the two B pins of MQ-2 is a DC signal and changes with the smoke concentration. Connect the B pin to the PA0 of the STM32 microcontroller [11].
- 3) Carbon monoxide will be produced in the early stage of a fire. After carbon monoxide enters the human body, it will combine with hemoglobin in the blood, causing hypoxia in the body tissue, causing the human body to faint and suffocate to death. Therefore, it is very important to transmit the concentration of carbon monoxide to the fire control center in real time to guide fire rescue.

B. NB IOT Communication Module

The system selects the BC35-G wireless communication module to send data and receive commands, which is a multi-band NB-IoT wireless communication module with very low power consumption, high sensitivity and low cost. And it supports multiple network protocol stacks with significant advantages in positioning, power consumption, data transmission rate and other module performance and system security [12]. The SIM card adopts the special NB-IOT network card provided by China Telecom to store temporary data [13], user information and encryption key. The application circuit is shown in Fig. 3.

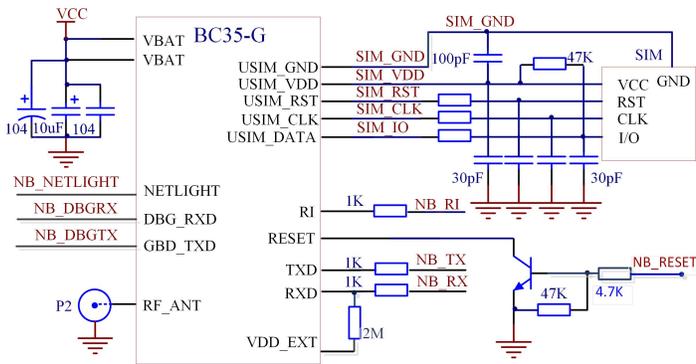


Fig. 3. Application Circuit.

C. Power Module Design

The external 3.7V lithium battery is chosen to power the whole hardware terminal, because the power supply voltage of each module is not consistent, the power supply voltage of the main control module is 3.3V, the power supply voltage of the data acquisition module is 5V, and the power supply voltage of the NB-IoT module is 3.7V. Therefore, the voltage needs to be converted into 3.3V and 5V to power the main control module and the data acquisition module respectively.

The lithium battery power supply module adopts the TP5410 chip for charging and boosting. The chip integrates the functions of charging and discharging, 5V constant voltage boosting, etc., and converts the lithium battery 3.7V voltage into 5V voltage through the boost converter. The circuit schematic diagram is shown in 3.3 shown. VUSB represents the voltage provided by the external device connected to the USB interface. The BAT pin is connected to a 3.7V lithium battery. The VOUT pin represents the circuit output voltage, which is converted to 5V by the booster.

RT8059 is selected as the 3.3V voltage regulator chip. This chip can effectively reduce the 5V voltage to 3.3V and conduct voltage stabilizing output. Its output current can reach 1 A. The power supply voltage regulator circuit is shown in Fig. 4.

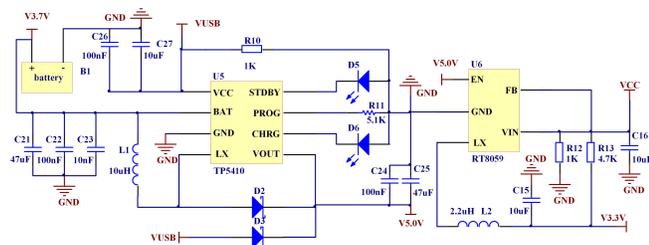


Fig. 4. Power Supply Voltage Regulator Circuit.

V. TERMINAL SOFTWARE DESIGN

The role of the sensing terminal software is to realize the collection of environmental information, data processing and complete data reporting. The terminal device software execution process is shown in Fig. 5. After the terminal device is powered on, the hardware enters the initialization state. After the initialization is completed, the BC35-G module

connects the device to the network, and checks the network attachment status by calling the AT command. When the network attachment is successful, it connects to the cloud platform for data transmission services. After the NB network is successfully attached, the sensing terminal (southbound device) is connected to the IoT cloud platform. After the connection is successful, the CDP server is configured. At this time, the device is in the wake-up state and remains connected to the cloud platform, and the device can communicate with the specified application server. The sensing terminal first collects sensor data, fuses the data, determines whether there is a gas leak or a fire, if so, sends an alarm message, and then sends the data to the cloud platform [14]. The main controller encodes the acquired data according to the defined binary format, and constructs the fused data into CoAP packets and sends them to the cloud platform. The cloud platform parses the CoAP message, decodes it by calling the decode interface, and converts it into a unified json data format to complete the data reporting. When the cloud platform issues a command, it calls the encode interface to encode and convert it into a binary code stream to construct a CoAP message and send it to the sensing terminal. The main controller parses the command and returns a response to control the operation of related external devices.

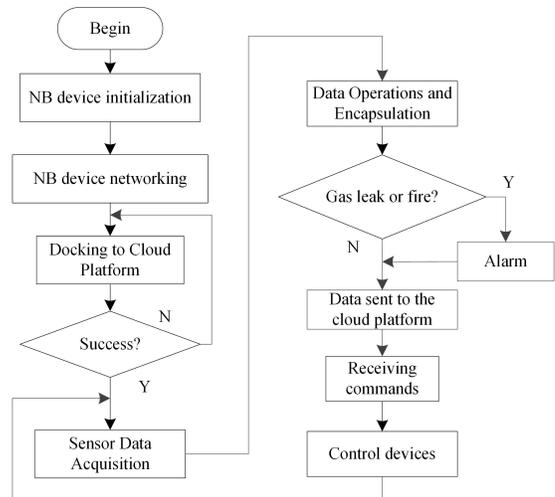


Fig. 5. End Device Software Execution Flow.

A. STM32 Project Configuration

The terminal software development uses the integrated development environment Keil uVision5. Since the STM32 requires a lot of initialization configuration during development, such as pin definition, clock configuration, etc. If the direct operation register development method is used, the process will be very cumbersome. In order to reduce the development time and energy, the STM32CubeMX software is used to configure the initialization code of the STM32 project in a graphical way. The generated STM32 initial project can be directly opened in Keil uVision5 and run.

Create a new project in STM32CubeMX, select the STM32F103C8 series MCU chip, configure the pin function according to the schematic diagram, and then configure the initialization parameters.

The main pin configuration functions are as follows:

- 1) PC14, PC15 pins: configured as RCC (clock system) clock source HSE (high-speed clock), and defined as Crystal/Ceramic Resonator (crystal/ceramic crystal) mode.
- 2) PD0, PD1 pins: configured as RCC clock source LSE (low speed clock), and defined as Disable mode.
- 3) PA13, PA14 pins: configured as program download pins and set to serial line mode.
- 4) PA2, PA3 pins: configured as USART2, which are serial communication pins between STM32 and NB-IoT module.
- 5) PB9, PB10 pins: configured as serial port USART1, the main function is to establish the communication serial port between STM32 and PC, and forward the communication content between NB-IoT module and STM32.
- 6) PA6 pin: configured as output pin GPIO_OUT to collect temperature and humidity values.
- 7) PA4 and PA5 pins: configured as analog-to-digital conversion pins ADC1_IN4 and ADC1_IN5, which are responsible for collecting smoke concentration and CO concentration respectively.

After the pin and initialization configuration is completed, the initialization project code is compiled and generated, which can be called directly when writing the program for each module.

B. Fire Discrimination Algorithm

The terminal nodes collect parameters such as temperature and humidity, smoke concentration, and CO concentration in real time, and then the collected data are homogenized and normalized, and then fuses the multi-sensor data through the BP neural network algorithm, and outputs it after decision analysis [15]. In the home environment, humidity has little effect on fire, so the effect of humidity is not considered in this paper for fire discrimination [16].

(1) Data preprocessing

Due to the existence of many interference noises in the external environment, the measured data are often inaccurate, which will cause great interference to the result judgment, and even misjudgment and omission may occur. In order to improve the accuracy of the data, the data needs to be preprocessed. Assuming that there are n different sensors in the system, their output vectors corresponding to the moment t can be summarized as $X(t) = (x_1(t), x_2(t), \dots, x_n(t))$, denoted:

$$\mu_i = \frac{1}{K} \sum_{i=1}^K x_i(t) \quad (1)$$

$$\sigma_i = \sqrt{\frac{1}{K} \sum_{i=1}^K (x_i(t) - \mu_i)^2} \quad (2)$$

$$f(x_i(t)) = \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left(\frac{-(x_i(t) - \mu_i)^2}{2\sigma_i^2}\right), i \in [1, n] \quad (3)$$

where K is the number of sample data, is the sample arithmetic mean, and is the sample standard deviation. And from equations (1) to (2), it can be inferred that the homogenization formula of heterogeneous sensor data is:

$$y_i(t) = \frac{|f(x_i(t)) - f(\mu_i)|}{f(\mu_i)} = \left| \exp\left(\frac{-(x_i(t) - \mu_i)^2}{2\sigma_i^2}\right) - 1 \right| \quad (4)$$

Thus, the input vector can be transformed into: $Y(t) = (y_1(t), y_2(t), \dots, y_n(t))$. If some of the data fluctuate greatly during the measurement process, it may happen that the data with smaller values in the output data is assimilated by the data with larger values. To avoid this event, this paper uses the normalization method to compress the data after homogenizing the data. The normalization formula is shown in (5).

$$x_i'(t) = \frac{y_i(t) - y_{\min}}{y_{\max} - y_{\min}} \quad (5)$$

where $y_i(t)$ is the input vector after homogenization, $x_i'(t)$ is the normalized covariate value, y_{\max} is the maximum input value, and y_{\min} is the minimum input value.

(2) Fire Data Identification

In this paper, the BP neural network algorithm is used to fuse the preprocessed data. BP neural network consists of input layer, implicit layer and output layer, and the most rapid descent method is used to learn the rules, and the actual result output is made infinitely close to the desired output by continuously adjusting the threshold and weights, so as to finally achieve the purpose of learning training.

The number of neurons in the input layer of the BP neural network is influenced by the dimension of the input data, and the number of neurons in the output layer is determined by the specific application, while the selection of neuron nodes in the hidden layer is particularly important, and the appropriate selection number will directly affect the performance of the BP neural network, which usually uses equation (6) to select the neuron nodes in the hidden layer.

$$l = \sqrt{m + n} + a \quad (6)$$

where m is the number of input layers, n is the number of output layers, $a \in [1, 10]$.

In order to avoid the problem of solving linear indistinguishability brought by linear mapping, the S-type activation function $f(x) = 1/(1 + e^{-x})$ is used in the BP neural network in this paper.

In this paper, the three parameters of temperature, smoke concentration, and CO concentration are used as the input layer of the BP neural network. And the open fire probability, shaded combustion probability, and no fire probability are used as the

feature outputs according to the type of fire occurrence. The training output value is $[0, 1]$ between. The number of input and output neuron nodes is 3. According to formula (6), the number of hidden layer neuron nodes is set to 10, and the structure of BP neural network is shown in Fig. 6.

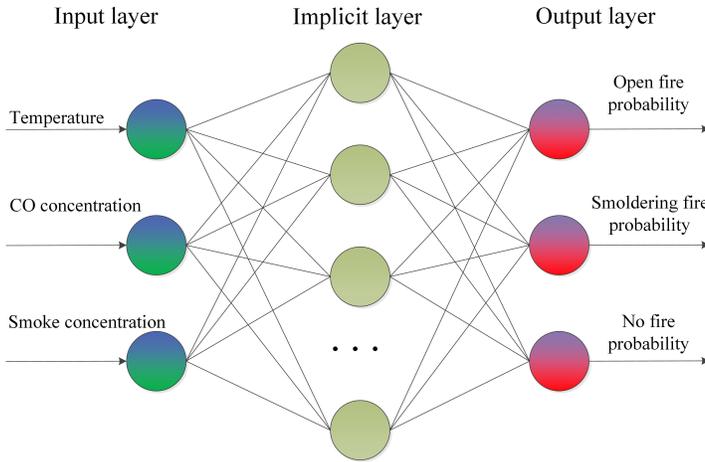


Fig. 6. Neural Network Structure.

The BP neural network structure is determined, and the network parameters are set:

Network inputs: $X_i = [x_1^i, x_2^i, x_3^i]$, group i temperature, CO concentration, smoke concentration;

Expected output: the expected output values of open fire probability, Smoldering fire probability and no fire probability in group i ;

Implicit layer input: $A_i = [a_1^i, a_2^i, a_3^i]$;

Implicit layer output: $B_i = [b_1^i, b_2^i, \dots, b_{10}^i]$;

Output layer input: $L_i = [l_1^i, l_2^i, \dots, l_{10}^i]$;

Output layer output: $C_i = [c_1^i, c_2^i, c_3^i]$, the actual output value of open fire probability, Smoldering fire probability and no fire probability in group i ;

The weight of the input layer and the implicit layer is w_{ij} , and the threshold is θ_j ; the weight of the hidden layer and the output layer is v_{jt} , and the threshold value is γ_t .

The network parameters are set and the feature-level fusion is performed. The core is iterative learning and training. The specific learning and training process is as follows:

- 1) Initialize the weights and thresholds, assign any real numbers within the interval $(-1, 1)$ of the weights and thresholds, and select the error function e , the calculation precision value ε and the limited number of learning times M .
- 2) Randomly select a group of samples k as input and target samples: the k th group of input $X_k = [x_1^k, x_2^k, x_3^k]$, the expected output $Y_k = [y_1^k, y_2^k, y_3^k]$.
- 3) Calculate the implicit layer inputs as:

$$A_i = \sum_{i=1}^3 (w_{ij}x_i - \theta_j) \quad (7)$$

The output is:

$$b_j = f(A_i) = \frac{1}{1 + e^{-A_i}}, j = 1, 2, \dots, 10 \quad (8)$$

- 4) Calculate the output layer input as:

$$L_j = \sum_{j=1}^{10} (v_{jt}b_j - \gamma_t) \quad (9)$$

The output is:

$$c_t = f(L_j) = \frac{1}{1 + e^{-L_j}}, t = 1, 2, 3 \quad (10)$$

- 5) Calculate the unit error between the actual output and the expected output:

$$d_t^k = (y_t^k - c_t)c_t(1 - c_t), t = 1, 2, 3 \quad (11)$$

- 6) Calculate the error of each unit in the middle layer as:

$$e_j^k = \left(\sum_{t=1}^3 d_t^k v_{jt} \right) b_j(1 - b_j), j = 1, 2, \dots, 10 \quad (12)$$

- 7) Output layer weights and threshold correction:

$$v_{jt}(N + 1) = v_{jt}(N) + \alpha d_t b_j \quad (13)$$

$$\gamma_j(N + 1) = \gamma_j(N) + \alpha d_t \quad (14)$$

where, N is the learning rate, $\alpha \in (0, 1)$.

- 8) Implicit layer weights and thresholds:

$$w_{ij}(N + 1) = w_{ij}(N) + \alpha [(1 - \eta)e_j x_i^k + \eta e_j x_1^{k-1}] \quad (15)$$

$$\theta_j(N + 1) = \theta_j(N) + \alpha [(1 - \eta)e_j^k + \eta e_j^{k-1}] \quad (16)$$

- 9) Continuously update the replacement weights and thresholds, and perform iterative training until the error meets the preset accuracy or reaches the limit of the number of learning times, and then the training can be ended.

In order to verify the accuracy of the system fire detection algorithm, the system measurement data, SH3 polyurethane plastic fire and SH6 wood fire are selected to form the original samples, and a total of 70 sample data are formed. Part of the original data and expected output are shown in Table I.

Homogenize and normalize the original data according to equations (1) to (5), and the data will be limited to the $[0, 1]$ interval after normalization. The processed sample data are shown in Table II.

Using matlab simulation software to simulate the neural network, 60 groups of sample data are selected from the existing test data for network training, of which the number of learning times is 50, the learning rate is 0.1, and the number of neurons in the hidden layer is 10. The neural network error obtained by BP neural network algorithm is shown in Fig. 7. Where, the abscissa is the number of training times, and the ordinate is the neural network error [17].

TABLE I. RAW DATA AND EXPECTED OUTPUT

Serial number	Input Value			Expected output		
	CO concentration	Smoke concentration	Temperature	Open fire probability	Smoldering probability	No fire probability
1	2.66	0.08	672	0.85	0.1	0.15
2	2.79	0.04	684	0.9	0.05	0.05
3	3.1	0.06	458	0.9	0.05	0.05
4	2.52	0.032	121	0.6	0.2	0.2
5	2.99	0.056	278	0.55	0.25	0.2
6	3	0.08	102	0.25	0.7	0.05
7	2.93	0.06	52	0.1	0.85	0.05
8	2.87	0.058	47	0.05	0.65	0.3
9	2.71	0.062	26	0.05	0.1	0.85

TABLE II. PREPROCESSED SAMPLE DATA

Serial number	Input Value			Expected output		
	CO concentration	Smoke concentration	Temperature	Open fire probability	Smoldering probability	No fire probability
1	0.24	1	0.98	0.85	0.1	0.15
2	0.47	0.17	1	0.9	0.05	0.05
3	1	0.58	0.66	0.9	0.05	0.05
4	0	0	0.14	0.6	0.2	0.2
5	0.81	0.5	0.38	0.55	0.25	0.2
6	0.83	1	0.12	0.25	0.7	0.05
7	0.71	0.58	0.04	0.1	0.85	0.05
8	0.6	0.54	0.03	0.05	0.65	0.3
9	0.33	0.63	0	0.05	0.1	0.85

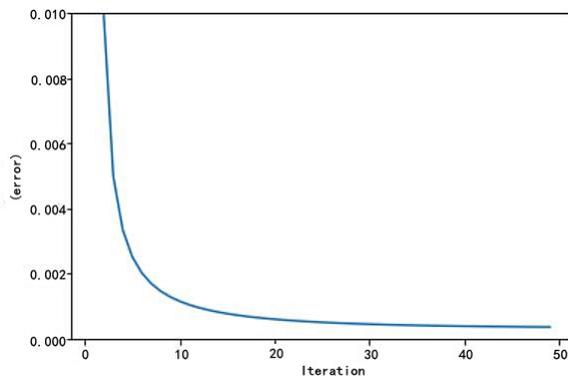


Fig. 7. Neural Network Error.

When the error is less than 0.0001, the training of the BP neural network is completed. The preprocessed sample data in Table II are tested experimentally, and the sample data are input into the trained BP neural network to obtain the actual output values of several different fire probabilities. The test results are shown in Table III.

Through the above 9 sets of test data, it is known that the absolute error of open flame probability is 0-0.024, and the average absolute error is 0.00567; the absolute error of smoldering probability is 0.002-0.035, and the average absolute error is 0.01067; the absolute error of no fire probability is 0.010-0.091, and the average absolute error is 0.02056, it can be seen that the trained neural network conforms to the actual output. The fire detection algorithm can improve the measurement accuracy of the fire detection device in practical applications, and can better reduce the probability of false alarms and missed alarms.

VI. IOT PLATFORM DESIGN

The IoT platform used in this system is Huawei's Ocean-Connect platform, and the design process is as follows [18].

- 1) Create a home fire system product application, record the application ID and key;
- 2) Write a profile file to define the capabilities and characteristics of the NB device;
- 3) Design a codec plug-in, parse the reported data and encode the issued commands;
- 4) Register NB-IoT device, connect the device to the network. After the device is connected to the network, the platform can receive device data to realize the connection and management of the NB-IoT module and the OceanConnect platform;
- 5) Debug, create online debugging and testing equipment, and use application simulator and NB equipment simulator to simulate the process of data reporting and command issuance.

The device Profile file developed in this paper mainly defines four attributes, smoke (smoke), CO concentration (CO), temperature (temp), humidity (humi), and also adds the down command field, sets CmdValue as the exhaust fan on command. Cmdvalue = 1 means to turn on one exhaust fan, cmdvalue = 0 means to turn off the exhaust fan, and the attribute type is int; Set alarm as the alarm command, take 1 as the alarm and 0 as the off alarm [19].

NB-IoT devices generally have higher requirements for power saving and use binary format. However, the IoT platform communicates with the application side using JSON format. Therefore, coding plug-ins need to be developed for the IoT platform to call in order to complete the conversion between binary format and JSON format. When the terminal reports

TABLE III. SAMPLE TEST DATA

Serial number	Input Value			Expected output		
	CO concentration	Smoke concentration	Temperature	Open fire probability	Smoldering probability	No fire probability
1	0.24	1	0.98	0.845	0.096	0.059
2	0.47	0.17	1	0.900	0.046	0.054
3	1	0.58	0.66	0.895	0.085	0.020
4	0	0	0.14	0.624	0.203	0.173
5	0.81	0.5	0.38	0.557	0.242	0.201
6	0.83	1	0.12	0.241	0.735	0.024
7	0.71	0.58	0.04	0.100	0.852	0.048
8	0.6	0.54	0.03	0.049	0.652	0.299
9	0.33	0.63	0	0.050	0.103	0.847

temperature, humidity, CO, and smoke data messages, the message name and data type must match the definitions of the corresponding fields in the profile, that is, they are consistent [20].

After the design of the codec plug-in is completed, you can add a real device or a new virtual device for debugging. After the debugging is passed, you can develop Web applications. This system uses the OceanBooster platform to develop northbound applications. OceanBooster supports forms, text, buttons, background pictures, etc. When adding new menus, external links can be added, with the ability to analyze device statistics, one-click device commands can be issued, and supports docking to third-party systems to quickly build WEB-side applications. According to the NB-IoT device developed by the profile file and codec above, add device monitoring components, button components, switch components, device status trend components, etc., set the style and layout of each component, and connect the components with the products developed in the IoT platform, connect the attributes, services, and commands of the components correspondingly, and realize the uploading and sending of commands and data. The interface of the monitoring system is shown in Fig. 8. The current room temperature is 18°C, humidity is 71%, smoke concentration is 1%, and CO concentration is 0; the devices registered in the IoT platform can be selected to display the current status of each device, and the audible and visual alarms and exhaust fans can be manually controlled [21].



Fig. 8. Monitoring System Interface.

VII. CONCLUSION

By analyzing the functional requirements of the terminal system, the design scheme is determined, and the hardware

and software of the terminal system are designed. The system hardware terminal composed of STM32F103C8T6 as the main controller, each sensor data acquisition module, BC35-G wireless communication module and other peripheral circuits is designed. Through the design of the system terminal software, the sensor data acquisition module can collect the indoor environmental parameters (temperature and humidity, smoke concentration, etc.) of home living in real time, and after being processed by the STM32 main controller, the terminal data is sent to the IoT cloud platform by the BC35-G networking. And the threshold method is used to realize the indoor gas detection and alarm function. The three raw data of temperature, CO concentration and smoke concentration are processed by homogenization and normalization methods, and the BP neural network is applied to data fusion of the preprocessed data. Through continuous iterative learning and training, the probability of fire occurrence is obtained, and accurate fire detection is realized. The main work of the paper is as follows:

- 1) The system architecture design is completed and the NB-IoT home fire monitoring terminal hardware and software function programs are designed. The terminal node can collect indoor environment data (temperature and humidity, smoke concentration, etc.) and detect fire in living environment in real time while setting the threshold value through sensor data to achieve gas leak detection; Use data fusion technology to realize fire detection function; use NB-IoT wireless transmission technology to regularly upload data to the IoT platform, which has the characteristics of long-distance transmission and strong practicability.
- 2) Based on Huawei's OceanConnect platform, the NB-IoT intelligent gateway is built. By defining Profile files and developing and deploying codec plug-ins, the data conversion function is realized, providing a communication bridge between the system terminal and the application side. At the same time, the CoAP protocol is used as the communication protocol between the hardware terminal and the IoT platform, which further reduces the power consumption of the hardware terminal.

From the experimental results, the system meets the overall requirements of the design. The home fire monitoring system designed by using the emerging NB-IoT technology solves the problems of high cost, high power consumption and short transmission distance of traditional Internet of Things technology, and has high practical value; the fire detection algorithm and BP neural network are used to train the sample model,

which improves the accuracy of fire detection and reduces the misjudgment rate; the Huawei OceanConnect platform is used to manage data and equipment, and to process and store data at the same time, reducing application costs and improving overall system performance.

In the follow-up research, the training samples will be further enriched, the humidity data will be incorporated into the network structure, and the neural network algorithm will be optimized so that it can be applied to more complex fire environments such as public buildings; a mobile APP application will be designed, and the monitoring system of this paper will be connected to the fire protection big data application platform of the government to realize grading early warning of police situation and rapid linkage response. With the rapid development of computer technology and network technology, these ideas are expected to be realized as soon as possible.

ACKNOWLEDGMENT

This work was financially supported by science foundation of the Department of Education of Jiangxi Province (GJJ150676).

REFERENCES

- [1] W. Teng, J. Chun, and H. Qiong, "Multi-band microstrip antenna applied to nb-iot / wlan / wimax," *Chinese Journal of Electron Device*, vol. 06, no. 42, pp. 1518–1521, 2019.
- [2] Z. Xiaoxin, W. Qichao, L. Feng, and Y. Baosheng, "Research on the application of narrow-band internet of things in power anti-stealing of special transformer user," *Chinese Journal of Electron Devices*, vol. 01, no. 44, pp. 178–181, 2021.
- [3] L. Dan, L. Xin, and L. Dichao, "Design and implementation of remote intelligent nursing system for the elderly," *Modern Computer*, vol. 35, no. 635, pp. 97–100, 2018.
- [4] E. L. S, L. J. A, and e. a. Moore A A, "Multi-band microstrip antenna applied to nb-iot / wlan / wimax," *Examining the effects of remote monitoring systems on activation, self-care, and quality of life in older patients with chronic heart failure*, vol. 01, no. 30, pp. 51–57, 2015.
- [5] F. Hui, CHENGTing, and CHENBao-guo, "Design of community fire alarm system based on zigbee," *Journal of Baicheng Normal University*, no. 43-48, 2021.
- [6] Q. Bin, H. Bing, and W. Xiao-juan, "Design of wireless fire alarm system based on lora and gprs," *Fire Science and Technology*, no. 242-245, 2021.
- [7] Z. Zhi-hua, X. Kai-li, and L. Zeng, "Study on the multi-sensor fire detection technology based on probabilistic neural network algorithm," *Fire Science and Technology*, no. 1404-1406, 2017.
- [8] X. Rongquan and X. Zhisheng, "Application of bp neural network on fire detection technology," *Journal of Railway Science and Engineering*, no. 140-145, 2014.
- [9] O. K. O, J. S. N, N.-O. E, and et al, "A wire-less sensor network based fire protection system with sms alerts," *International Journal of Mechanical Engineering and Technology(IJMET)*, vol. 2, no. 10, pp. 44–52, 2019.
- [10] B. Cheng, H. Kuihong, W. Shixi, and H. Tianfeng, "Circuit design of temperature control module for large passenger aircraft fire monitoring laser," *Instrument Technique and Sensor*, vol. 12, no. 443, pp. 45–48, 2019.
- [11] L. Bo and Y. Lipeng, "Design of smart home system based on stm3," *Electronic Design Engineering*, vol. 07, no. 28, pp. 177–180, 2020.
- [12] L. Chuncheng, Y. Yun, C. Liang, H. Jian, and C. Xiujian, "Self-powered low-power wireless forest fire monitoring system," *Chinese Journal of Electron Devices*, vol. 03, no. 44, pp. 707–712, 2021.
- [13] Y. Weizhong, W. Yachun, Y. Yao, and S. Jingbo, "Soil moisture monitoring system based on narrow band internet of things," *Journal of Agricultural Machinery*, vol. 00, no. 201, pp. 243–247, 2019.
- [14] M. Aliff, N. Samsiah, M. I. Yusof, and A. Zainal, "Development of fire fighting robot (qrob)," *International Journal of Advanced Computer Science and Applications*, vol. 10, no. 1, pp. 147–142, 2019.
- [15] O. Chamorro-Atalaya, D. Arce-Santillan, G. Morales-Romero, A. Quispe-Andia, N. Trinidad-Loli, E. Auqui-Ramos, C. Leon-Velarde, and E. Gutierrez-Zubieta, "Level transducer circuit implemented by ultrasonic sensor and controlled with arduino nano for its application in a water tank of a fire system," *International Journal of Advanced Computer Science and Applications*, vol. 12, no. 10, pp. 464–471, 2021.
- [16] K. Muhammad, S. Khan, M. Elhoseny, S. Hassan Ahmed, and S. Wook Baik, "Efficient fire detection for uncertain surveillance environment," *IEEE Transactions on Industrial Informatics*, vol. 15, no. 05, pp. 3113–3122, 2019.
- [17] D. Yongmei, S. Huixia, Z. Ling, S. Lei, C. Rijun, and H. Junhui, "Design of fire monitoring system based on matlab," *Electronic Component and Information Technology*, vol. 01, no. 05, pp. 58–61, 2021.
- [18] A. Khan, F. Bibi, M. Dilshad, S. Ahmed, Z. Ullah, and H. Ali, "Accident detection and smart rescue system using android smartphone with real-time location tracking," *International Journal of Advanced Computer Science and Applicatio*, vol. 09, no. 06, pp. 341–355, 2018.
- [19] W. Xiaoyan and D. Qisheng, "Design of fire monitoring system based on wsn and labview," *Instrument Technique and Sensor*, vol. 07, no. 354, pp. 35–37+41, 2012.
- [20] H. Juan, C. Haodong, L. Yongfeng, and L. Qince, "Design of zigbee and android based fire monitoring system for ancient buildings," *Fire Science and Technology*, vol. 07, no. 38, pp. 973–976, 2019.
- [21] L. Barik, "Iot based temperature and humidity controlling using arduino and raspberry pi," *International Journal of Advanced Computer Science and Applications*, vol. 10, no. 9, pp. 494–502, 2019.

The Effect of Natural Language Processing on the Analysis of Unstructured Text: A Systematic Review

Walter Luis Roldan-Baluis, Noel Alcas Zapata, Maria Soledad Mañaccasa Vásquez

Postgraduate School, Doctorate in Education
Cesar Vallejo University, Lima, Perú

Abstract—The analysis of the unstructured text has become a challenge for the community dedicated to natural language processing (NLP) and Machine Learning (ML). This paper aims to describe the potential of the most used NLP techniques and ML algorithms to address various problems afflicting our society. Several original articles were reviewed and published in SCOPUS during 2021. The applied approach was retrospective, transversal and descriptive. The data collected were entered into the SPSS statistical software v25 and among the findings, it was determined that the most used NLP technique was the Term frequency - Inverse document frequency (TF-IDF), while the most used supervised learning algorithm was the Support Vector Machines (SVM). Likewise, the predominant deep learning algorithm was Long Short-Term Memory (LSTM). This research aims to support experts and those starting in research to identify the most used algorithms of NLP and ML.

Keywords—Artificial intelligence; natural language processing; machine learning; unstructured text analysis

I. INTRODUCTION

The Internet has become an exclusive ally for any institution. According [1], there were more than 5'168,000,000 users worldwide, of which Asia accounts for 53.4%, and ranks first. Latin America and the Caribbean are positioned in ranks fourth with 9.6%. According to [2], in Spain the number of users reached 91% of the population. The author in [3] points out that there are currently more than 1'900,000,000 web pages. It also points out that 167 million videos are generated in a minute on the Tik Tok platform; likewise, Amazon customers invest USD 283,000 in e-commerce. It is concluded that in every second large amounts of data are produced in various formats such as images, audios, videos and texts.

The massive amount of data implies the need to automate human tasks through the fast advance of technological innovation. It can be used for decision making in an efficient and effective way. According to [4], such innovation includes Artificial Intelligence (AI). The author in [5] points out that AI trains computers to learn from experience and to do the work of human beings. This field has had an intensified growth due to the COVID-19 pandemic. In the research [6] states that AI surpasses the cognitive abilities of man. AI is an interdisciplinary field [7], of computing capable of solving problems of medicine, psychology, education, health, information technologies - TIC, among others.

On the Internet platform, there is a lot of traffic, users are producing a high volume of unstructured texts; it is difficult to determine which websites are visited by users. The author in

[8] propose a model made based on Natural Language Processing (NLP) techniques and neural networks to identify the websites visited by users by translating this problem into a text classification context. This solution is advantageous for the digital marketing because it allows the loyalty of users.

Social networks are platforms on which there are a high proliferation of comments, with absolute freedom and without restrictions from attacks, insults, discriminatory speeches, hatreds and other offensive terms. In the research work [9] proposes a text classification model to detect cyberbullying consisting of a neural network framework that examines the content of the text in order to analyze the effect of the extracted characteristics. The usefulness of this study lies in identifying solid mechanisms for the detection of cyberbullying.

Regarding the scope of the research, systematic review articles published in the SCOPUS Database, period 2021, were reviewed. For example, [10] submitted a systematic review article to provide evidence on the properties of text data used to train machine learning approaches and how they can be applied in clinical practice. In another review article, [11] highlighted the usefulness of NLP and Machine Learning (ML) to structure the comments of free texts issued by patients of health organizations. This led to identify that there is no systematic research that describes the frequency of the ML algorithms used in the various original articles. This work aims to fill this gap, being this the main motivation to carry out this research.

The objective of this study is to systematically review the bibliography of the application of natural language processing to analyze, interpret and classify the high production of unstructured texts produced in digital format. Also, to describe the frequency of ML algorithms such as supervised, unsupervised and deep learning. In this context, the aim is to answer the questions raised in Table I. Question one aims to identify NLP application fields. These features include text preprocessing techniques such as tokenization, etc. Question two describes the frequency of ML algorithms for data analysis. Finally, question three refers to the frequency of deep learning algorithms; this question has been given preference since algorithms are very specialized.

TABLE I. RESEARCH QUESTIONS

No.	Question
1	What are NLP and ML application fields?
2	What is the frequency of ML algorithms for text analysis?
3	What is the frequency of DL algorithms?

II. LITERATURE REVIEW

Natural Language Processing – NLP is a branch of artificial intelligence and a resource to carry out qualitative tasks of unstructured information, based on mathematical and statistical algorithms on large amounts of data. In this regard, [11] pointed out that the NLP is a computer analytical technique used to extract information from an unstructured text into a structured form, for which syntactic processing of a text is done; it also captures the meaning and identifies links based on semantic relationships. The author in [12] indicates that NLP is a technique of automatic extraction of information from different electronically written resources at the level of documents, words, grammar, meaning, and context. Likewise, [13] stated that the NLP is a key tool for information automation and extraction that can process large amounts of data and its application is useful for issuing reports from the radiology area of a hospital.

Machine Learning (ML) is used for creating models that allow computers to learn without being programmed. In this regard, [11] affirm that ML is a set of statistical algorithms that can train and test a group of data to detect patterns, predict feelings within a text. In the research [14] point out that ML is the process that detects and exploits patterns and trends that are "hidden" in the production of unstructured texts. The author in [15] indicate that ML is a set of machine learning algorithms built into machines to provide knowledge about processes quickly and efficiently. ML is classified into three fields, supervised learning (S), unsupervised learning (US), and reinforcement learning. The present work contemplates the use of algorithms of the first two fields mentioned.

Supervised learning uses algorithms that learn iteratively from data. They find hidden information by which computers learn. The author in [11] point out that algorithms try to predict and classify texts. For example, in the electronic documentation of a health service, the algorithms are able to identify the most common issues expressed by patients. Likewise, [16] points out that supervised ML divides the input data set into training and testing. The training data set has an output variable that must be predicted or classified.

Unsupervised learning uses algorithms to identify patterns and detect anomalies such as fraud, scam of potential users, among others. In this regard, [11] indicates that it is a technique that identifies models or patterns of behavior without the need to know the target attribute or objective that could be present in a text. In the research work [16] points out that the algorithms learn some characteristics from the data. One of the best-known models is the clustering.

The NLP is taking a lot of relevance in the sentiment analysis (SA), positive or negative, in the analysis of unstructured text; a source of application is the comments that are made on social networks. In that aspect, [17], making use of the tasks of NLP and ML methods, propose a model of word processing for SA that uses the comments made on Twitter. The first phase consists of collecting the text, cleaning it, preprocessing, extracting features from a text and then categorizing the data. The proposed corpus is multidisciplinary and can be used in the area of market analysis, customer behavior, survey analysis, and brand monitoring, among others.

This contribution is used as a basis for broadening the range of real applications.

The usefulness of NLP and ML has a high level of application in the medicine field. It can be applied to determine the misuse and abuse of prescription drugs in comments made on social networks. In this regard, [18] propose a model to detect self-reports of prescription drug abuse from Twitter. Using these public data, it develops a continuous monitoring system to classify the class of "abuse or misuse".

III. MATERIAL AND METHOD

A. Introduction

The PRISMA method is a structured tool with a systemic approach that helps to present the results of a research. According to [19], the Preferred Reporting Items for Systematic Reviews and Meta-Analyses – PRISMA 2020 is conceptualized as a series of recommendations that contribute to selecting, evaluating and synthesizing for better clarity and transparency of research. In fact, [20], [21] point out that the PRISMA declaration is an essential strategy for conducting good research and publishing the results. In the area of objectivity, this research has been divided into four phases, according to the process proposed by:

- 1) Retrieval of publications.
- 2) Review of titles and abstracts.
- 3) Revision of the full text.
- 4) System information collection.

With regard to the initial recovery phase, it is necessary to use a strategy that would allow efficient document searches. In this respect, [22] point out that the PICO strategy is relevant for raising research questions in order to optimize the placement of articles. The PICO system is an acronym and a component structure. According to [23], this format has four elements: problem, intervention, comparison and outcome. Table II shows the optimal search of documents, this strategy was adapted to the acronym PIO. In addition, the thesaurus Computer Classification System – ACM was used to identify the appropriate synonyms; the link is: <https://bit.ly/3dphAJP>.

From phase two: review, titles and abstracts; articles were located to be contrasted with the inclusion and exclusion criteria. Titles and abstracts were reviewed, then the method and results, in order to establish the search formula. The database consulted was Scopus, period 2021. In the third phase, the combination of keywords and synonyms was used with emphasis on the variables Natural Language Processing and text analysis. The logical operators AND and OR were used repeatedly until the appropriate formula was obtained. Table III shows the restricted query.

TABLE II. KEYWORDS AND SYNONYMS FOR THE PIO METHOD SEARCH

P	I	O
Natural Language Processing Natural language process Natural Language Text Computational linguistics Word processing NLP	text analysis text analytics text data	Corpus Classification

TABLE III. SEARCH CRITERIA FOR ORIGINAL SCIENTIFIC ARTICLES

Database	Search date	Search string
Scopus	December 15, 2021	OA(all) AND (TITLE-ABS("Natural Language Processing") OR TITLE-ABS("Natural Language Process") OR TITLE-ABS("Natural Language Text") OR TITLE-ABS("Computational linguistics") OR TITLE-ABS("Word processing") OR TITLE-ABS("NLP")) AND (TITLE-ABS("text analysis") OR TITLE-ABS("text analytics") OR TITLE-ABS("text data") OR TITLE-ABS("text classification") OR TITLE-ABS("Data extraction")) AND PUBYEAR > 2020 AND DOCTYPE(AR)

B. Selection of Criteria

Inclusion and exclusion criteria for the efficient search of research articles were identified in the PICO strategy. The query was held on December 19, 2021. The search was restricted since 2021 and 144 articles were located in Scopus database. To ensure the rigor and credibility of the selected articles, they were evaluated by extrapolating the criteria defined in Table IV.

TABLE IV. INCLUSION AND EXCLUSION CRITERIA USING THE PICOS MODEL

PICOS	Inclusion criteria	Exclusion criteria
Problem	Natural language processing – NLP in the text analysis	Natural language processing in formats other than texts (e.g., video, audio).
Intervention	NLP interventions in the data extraction and summaries of text analysis with free software (R language, Python)	NLP interventions in which actual text, using an NLP process, is not processed. Data extraction with licensed software. Chatbot.
Comparison	Comparison with other type of intervention such as the elaboration of the linguistic corpus.	Studies that have no other type of comparison.
Outcomes	Report on the impact of the intervention.	It does not contain a report on the impact of the intervention.
Study Type	Quantitative, qualitative, and mixed method studies of original articles.	Systematic review articles, meta-analysis, literature reviews, conferences, dissertations, protocol works, tutorials. Studies not conducted in English. Duplicate jobs and not available in full text.

IV. RESULTS

A. Search Results

A total of 144 articles were collected during the search process. 11 were deleted after reviewing the title and abstract (n=133) of each document. Then, the method and conclusions were reviewed with emphasis and those that did not meet the inclusion criteria were discarded (n=87). Finally, there were 46 potential articles for systematic review. Fig. 1 shows the flowchart of the search strategy.

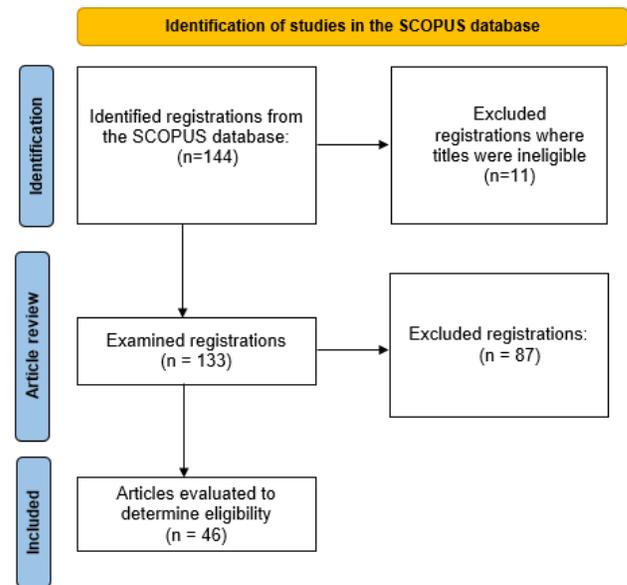


Fig. 1. Flowchart of the Literature Review Process.

B. Description of Included Studies

The application fields or sectors that have benefited from the NLP and ML application correspond to the domains such as aviation, medicine, cyberbullying, education, engineering, technology, among others. In this regard, the medical sector has benefited from 13 studies, 28.76%. The education system from 10 studies, 21.74%. The Technology field has seven articles, 15.22%, among others. Table V shows the details.

TABLE V. FIELDS OF APPLICATION AND FREQUENCY OF ARTICLES

Application field	Frequency	%	Author(s)
Aviation	1	2.17	[24]
Natural disaster relief	1	2.17	[25]
Cyberbullying	1	2.17	[9]
Construction	1	2.17	[26]–[27]
Software Development	3	6.52	[28]–[30]
Education	10	21.74	[31]–[40]
Finance	1	2.17	[41]
Engineering	1	2.17	[42]
Marketing	2	4.35	[8]–[17]
Medicine	13	28.26	[43]–[55]
Business organization	1	2.17	[56]
Politics	1	2.17	[57]
Psychology	1	2.17	[58]
Information security	1	2.17	[59]
Technology	7	15.22	[60]–[66]
Urban transport	1	2.17	[67]

C. Frequency of NLP Algorithms

A number of NLP techniques were applied prior to the use of ML algorithms, and these NLP techniques are described in Appendix A. NLP techniques are associated with various algorithms, which are defined in Appendix B. After registering the data in the SPSS v25 statistical software, it has been discovered that the Term frequency - inverse document frequency (TF-IDF) algorithm was present in 22 articles, 47.82%. The Word2Vec algorithm was used 15 times, 32.60%. Glove algorithm at 14, 30.43%, while Bag of words (BoW) was used in 11 studies, 23.91%, and N-Gram was used in six articles, 13.04%.

On the other hand, seven studies, 15.21%, used three algorithms at the same time. Likewise, 12 articles, 26.09%, used two algorithms at the same time, while 26 articles used a single algorithm, 58.69%, in their research. The details of the NLP algorithms used are detailed in Appendix C.

D. Frequency of ML Algorithms

ML algorithms analyzed in this study are defined in Appendix D and grouped under supervised learning, unsupervised learning, and Deep Learning.

1) *Frequency of supervised learning algorithms (S)*: The Support Vector Machine (SVM) algorithm was used in 17 studies. The Naive Bayes (NB) algorithm was applied in 15 studies. The Random Forest (RF) algorithm has 10 studies. R has 9 studies. K-NN has 8 studies. RF has 5 studies. The Passive aggressive (PA) algorithm was used in two studies. The AdaBoost (ADA) algorithm has 1 study like Singular Value Decomposition (SVD) algorithm, Fig. 2 shows the details.

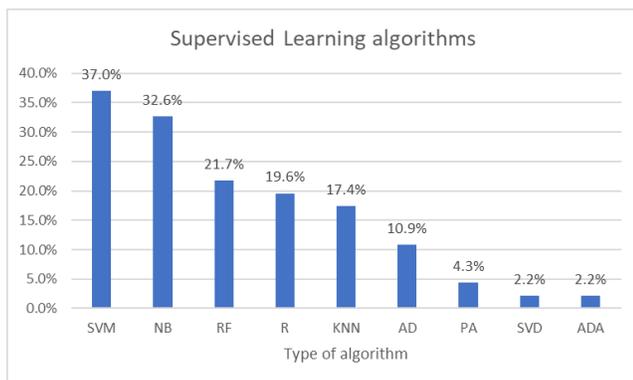


Fig. 2. Studies using S Algorithms.

2) *Frequency of unsupervised learning algorithms (US)*: The Latent Dirichlet Allocation (LDA) algorithm was used in three studies, 6.5%, while K-Means algorithm was used only in one study, 2.2%.

3) *Frequency of deep learning algorithms (DL)*: The Long Short Term Memory (LSTM) deep learning algorithm has 24 studies. Then, the Convolutional neural networks (CNN) algorithm has 12 studies. The Recurrent Neural Networks (RNN) algorithm has 9 studies. The Multilayer Perceptron (MLP) algorithm has 5 studies. The least used algorithms were

Gating Circulation Unit (GRU) algorithm with four studies and Artificial neural network (ANN) with two studies. Fig. 3 shows the details.

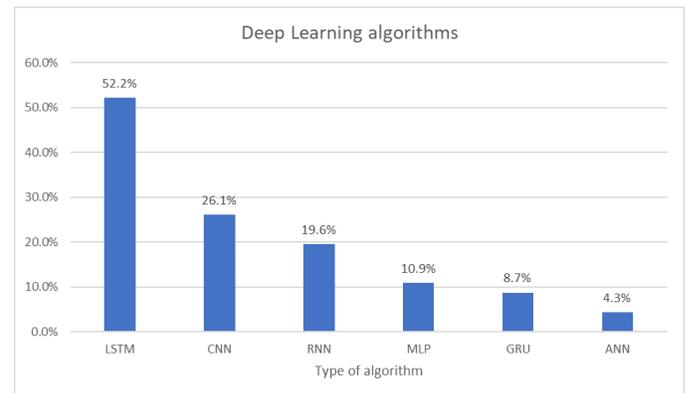


Fig. 3. Studies using DL Algorithms.

4) *Studies with hybrid algorithms, Supervised (S), Unsupervised (US) and Deep Learning (DL)*: Out of 46 articles, 10 (21.74%) use only S algorithms. In this regard, [24] and [52] use SVM. The author in [35] use NB algorithm. On the other hand, the study of [51] uses the LDA algorithm.

Three studies use S and US algorithms at the same time: [17] use two S algorithms: SVM, NB and 1 NS: K-Means. [67] use five S algorithms: SVM, NB, Regression (R), RF, KNN, and one US algorithm: LDA.

The author in [25] uses three algorithms: A supervised learning algorithm, the LDA, and two deep learning algorithms, GRU and CNN. Twenty-one studies, 45.65%, used only DL algorithms, in particular [8], [32], [48] y [68].

Eleven studies, 23.91%, used S and DL algorithms at the same time, in particular [9], [36], [38], [46], [47], [49], [53], [55], [57], [58] y [60].

The details of the ML algorithms used by the 46 studies can be found in Appendix C.

V. DISCUSSION

The popularity and proliferation of platforms working on the Internet such as web portals, social networks, and all digital media have created a massive social interaction between users, even more so because of the global COVID-19 pandemic that has led to the unprecedented increase in online learning and its consequent exponential production of unstructured texts. This phenomenon, according to [32], is allowing the increasing use of the NLP and ML field in text analysis for an efficient solution of real problems.

What are NLP and ML application fields?

It was discovered that sectors such as the health system, education, technology, engineering, software development, aviation, natural disasters relief, cyberbullying, construction, finance, marketing, politics, business organization, information security, psychology, and urban transport, benefit most. This reflects that NLP and ML can be applied to solve problems in any sector.

What is the frequency of ML algorithms for text analysis?

The analysis of the articles indicates that NLP preprocessing techniques such as tokenization, normalization, elimination of irrelevant words are necessary to apply ML algorithms, which allow having a positive impact to achieve the Garg & Sharma study objective [17]. TF-IDF, word2Vec, and Glove are among the most used NLP algorithms. The ML algorithms of supervised learning were: SVM, NB and RF. The least used algorithms were: PA, ADA and SVD. With respect to unsupervised learning algorithms, these were the least used. Only three studies used the LDA algorithm.

What is the frequency of DL algorithms?

With regard to Deep learning, the most used algorithm was LSTM with 22 articles, and the least used was ANN with only 2 articles. This approach becomes a primary tool for the NLP. However, it should be noted that ML algorithms can lead to error bias because it depends on the quality of data with which the research is carried out and especially on access to data since many institutions, unfortunately, restrict them, for example, hospitals [69].

Considering the works obtained, it can be said that the most used NLP technique was the TF-IDF. The most used supervised learning algorithm was the SVM, and with respect to neural networks or deep learning, it was the LSTM. On the other hand, according to [69], one of the main obstacles to applying the NLP and ML algorithms is access to data, representing a challenge for the AI community in reversing this situation.

VI. CONCLUSION

The most commonly used supervised learning algorithms for text analysis in the field of research are TF-IDF, word2Vec and Glove, while predominant deep learning algorithms are LSTM and ANN. In addition, this article complements the various studies regarding systematic reviews on NLP and ML, by describing the frequencies of influential algorithms and it is expected that this work will lead to further research to increase the cases of application of PLN and ML for the benefit of various fields such as health, education, transport, technology and others. Finally, it should be noted that improving the cognitive aspect of this science requires further research taking into account that the PLN and ML algorithms are universal, characteristic of mathematics and statistics.

REFERENCES

- [1] Internet World Stats, Global Threat Report 2021, 1d. C., 2022, <https://www.internetworldstats.com/stats.htm>.
- [2] S. Galeano, M4rketng Ecommerce, Qué pasa en Internet en un minuto en 2021, 2022, <https://marketing4ecommerce.net/que-pasa-en-internet-en-un-minuto-infografia/>.
- [3] Internet live stat, Total number of Websites, 2021. <https://www.internetlivestats.com/total-number-of-websites/>.
- [4] D. Gruson, «Big Data , inteligencia artificial y medicina de laboratorio : la hora de la integración», *Adv Lab Med*, 2(1), 5-7, 2021, <https://doi.org/10.1515/almed-2021-0014>.
- [5] H.P. Winston, Artificial Intelligence, Third, Enited States of America, 1992. <https://courses.csail.mit.edu/6.034f/ai3/rest.pdf>.
- [6] D.F. Arbeláez-Campillo, J.J. EspinozaV illasmil, M.J. Rojas-Bahamón, «Inteligencia artificial y condición humana: ¿Entidades contrapuestas o

- fuerzas complementarias?», *Revista de ciencias sociales*, 27(2), 502-513, 2021, doi:10.31876/rcs.v27i2.35937.
- [7] D. Garabato, Análisis no supervisado de observaciones atípicas en la misión espacial Gaia: optimización mediante procesamiento distribuido e integración en Apsis, Universidade da Coruña, 2020. <https://ruc.udc.es/dspace/handle/2183/26479>.
- [8] D. Perdices, J. Ramos, J.L. García-Dorado, I. González, J.E. López de Vergara, «Natural language processing for web browsing analytics: Challenges, lessons learned, and opportunities», *Computer Networks*, 198, 2-14, 2021, doi:10.1016/j.comnet.2021.108357.
- [9] C. Raj, A. Agarwal, G. Bharathy, B. Narayan, M. Prasad, «Cyberbullying detection: Hybrid models based on machine learning and natural language processing techniques», *Electronics (Switzerland)*, 10(22), 1-20, 2021, doi:10.3390/electronics10222810.
- [10] I. Spasic, G. Nenadic, «Clinical text data in machine learning: Systematic review», *JMIR Medical Informatics*, 8(3), 2020, doi:10.2196/17984.
- [11] M. Khanbhai, P. Anyadi, J. Symons, K. Flott, A. Darzi, E. Mayer, «Applying natural language processing and machine learning techniques to patient experience feedback: A systematic review», *BMJ Health and Care Informatics*, 28(1), 1-12, 2021, 10.1136/bmjhci-2020-100262.
- [12] S.R. Jonnalagadda, P. Goyal, M.D. Huffman, «Automating data extraction in systematic reviews: A systematic review», *Systematic Reviews*, 4(1), 2015, doi:10.1186/s13643-015-0066-7.
- [13] E. Pons, M. Loes, M. Braun, M. Hunink, J. Kors, «natural Language Processing in Radiology: A Systematic Review1», *Radiology*, 279(2), 329-343, 2021, doi:10.1148/radiol.16142770.
- [14] M. Ceriotti, C. Clementi, O. Anatole Von Lilienfeld, «Introduction: Machine Learning at the Atomic Scale», *Chemical Reviews*, 121(16), 9719-9721, 2021, doi:10.1021/acs.chemrev.1c00598.
- [15] M. Alloghani, D. Al-Jumeily, J. Mustafina, A. Hussain, A. Aljaaf, Springer Link, A Systematic Review on Supervised and Unsupervised Machine Learning Algorithms for Data Science, 2019.
- [16] M. Batta, «Machine Learning Algorithms - A Review», *International Journal of Science and Research (IJ)*, 9(1), 381-386, 2020, doi:10.21275/ART20203995.
- [17] N. Garg, K. Sharma, «Text pre-processing of multilingual for sentiment analysis based on social network data», *International Journal of Electrical and Computer Engineering*, 12(1), 776-784, 2022, doi:10.11591/ijece.v12i1.pp776-784.
- [18] M.A. Al-Garadi, Y.C. Yang, H. Cai, Y. Ruan, K. O'Connor, G.H. Graciela, J. Perrone, A. Sarker, «Text classification models for the automatic detection of nonmedical prescription medication use from social media», *BMC Medical Informatics and Decision Making*, 21(1), 1-13, 2021, doi:10.1186/s12911-021-01394-0.
- [19] D. Moher, «Reporting guidelines: Doing better for readers», *BMC Medicine*, 16(1), 18-20, 2018, doi:10.1186/s12916-018-1226-0.
- [20] R. Sarkis-Onofre, F. Catalá-López, E. Aromataris, C. Lockwood, «How to properly use the PRISMA Statement», *Systematic Reviews*, 10(1), 13-15, 2021, doi:10.1186/s13643-021-01671-z.
- [21] K. Kreimeyer, M. Foster, A. Pandey, N. Arya, G. Halford, S.F. Jones, R. Forshee, M. Walderhaug, T. Botsis, «Natural language processing systems for capturing and standardizing unstructured clinical information: A systematic review», *Journal of Biomedical Informatics*, 73, 14-29, 2017, doi:10.1016/j.jbi.2017.07.012.
- [22] E. Landa-Ramírez, A. de J. Arredondo-Pantaleón, «Herramienta pico para la formulación y búsqueda de preguntas clínicamente relevantes en la psicooncología basada en la evidencia», *Psicooncología*, 11(2-3), 250-270, 2014, doi:10.5209/rev_PSIC.2014.v11.n2-3.47387.
- [23] A. Pérez Ortiz, M. Ortega Luyando, A. Amaya Hernández, «Programas de prevención de obesidad infantil en México: una revisión sistemática PICO», *Psicología y Salud*, 31(2), 169-177, 2021, doi:10.25009/pys.v31i2.2686.
- [24] T. Madeira, R. Melício, D. Valério, L. Santos, «Machine learning and natural language processing for prediction of human factors in aviation incident reports», *Aerospace*, 8(2), 1-18, 2021, doi:10.3390/aerospace8020047.

- [25] S. Khatoun, M.A. Alshamari, A. Asif, M.M. Hasan, S. Abdou, K.M. Elsayed, M. Rashwan, «Development of social media analytics system for emergency event detection and crisismanagement», *Computers, Materials and Continua*, 68(3), 3079-3100, 2021, doi:10.32604/cmc.2021.017371.
- [26] M.L. Yu, M.H. Tsai, «ACS: Construction data auto-correction system-Taiwan public construction data example», *Sustainability (Switzerland)*, 13(1), 1-21, 2021, doi:10.3390/su13010362.
- [27] S. Moon, G. Lee, S. Chi, H. Oh, «Automated Construction Specification Review with Named Entity Recognition Using Natural Language Processing», *Journal of Construction Engineering and Management*, 147(1), 1-12, 2021, doi:10.1061/(asce)co.1943-7862.0001953.
- [28] Ş. Ozan, «Case studies on using natural language processing techniques in customer relationship management software», *Journal of Intelligent Information Systems*, 56(2), 233-253, 2021, doi:10.1007/s10844-020-00619-4.
- [29] M. Alenezi, Z. Mohammed, Y. Javed, «Efficient deep features learning for vulnerability detection using character ngram embedding», *Jordanian Journal of Computers and Information Technology*, 7(1), 25-38, 2021, doi:jjcit.71-1597824949.
- [30] P. Rani, S. Panichella, M. Leuenberger, A. Di-Sorbo, O. Nierstrasz, «How to identify class comment types? A multi-language approach for class comment classification», *Journal of Systems and Software*, 181, 2-17, 2021, doi:10.1016/j.jss.2021.111047.
- [31] L. Burdick, J.K. Kummerfeld, R. Mihalcea, «To batch or not to batch? Comparing batching and curriculum learning strategies across tasks and datasets», *Mathematics*, 9(18), 2021, doi:10.3390/math9182234.
- [32] M. Mujahid, E. Lee, F. Rustam, P.B. Washington, S. Ullah, A.A. Reshi, I. Ashraf, «Sentiment Analysis and Topic Modeling on Tweets about Online Education during COVID-19», *Applied Sciences*, 11(18), 1-25, 2021, doi:10.3390/app11188438.
- [33] R. Adipradana, B.P. Nayoga, R. Suryadi, D. Suhartono, «Hoax analyzer for Indonesian news using RNNs with fasttext and glove embeddings», *Bulletin of Electrical Engineering and Informatics*, 10(4), 2130-2136, 2021, doi:10.11591/eei.v10i4.2956.
- [34] O.C. Stringham, S. Moncayo, K.G.W. Hill, A. Toomes, L. Mitchell, J. V. Ross, P. Cassey, «Text classification to streamline online wildlife trade analyses», *PLOS ONE*, 16(7), e0254007, 2021, doi:10.1371/journal.pone.0254007.
- [35] S.G. Chethan, S. Vinay, «Analytical Framework for Binarized Response for Enhancing Knowledge Delivery System», *International Journal of Advanced Computer Science and Applications*, 12(10), 348-358, 2021, doi:10.14569/ijacsa.2021.0121157.
- [36] R.A. Farouk, M.H. Khafagy, M. Ali, K. Munir, R. M.Badry, «Arabic Semantic Similarity Approach for Farmers' Complaints», *International Journal of Advanced Computer Science and Applications*, 12(10), 348-358, 2021, doi:10.14569/IJACSA.2021.0121038.
- [37] C.X. Zhang, R. Liu, X.Y. Gao, B. Yu, «Graph Convolutional Network for Word Sense Disambiguation», *Discrete Dynamics in Nature and Society*, 2021, 1-12, 2021, doi:10.1155/2021/2822126.
- [38] J.L. Huan, A.A. Sekh, C. Quek, D.K. Prasad, «Emotionally charged text classification with deep learning and sentiment semantic», *Neural Computing and Applications*, 1, 1-11, 2021, doi:10.1007/s00521-021-06542-1.
- [39] H.X. Huynh, L.X. Dang, N. Duong-Trung, C.T. Phan, «Vietnamese Short Text Classification via Distributed Computation», *International Journal of Advanced Computer Science and Applications*, 12(7), 23-31, 2021, doi:10.14569/IJACSA.2021.0120703.
- [40] Y. Hu, H. Shen, W. Liu, F. Min, X. Qiao, K. Jin, «A Graph Convolutional Network with Multiple Dependency Representations for Relation Extraction», *IEEE Access*, 9, 1-14, 2021, doi:10.1109/ACCESS.2021.3086480.
- [41] D. Alsaleh, S. Larabi-Marie-Sainte, «Arabic Text Classification Using Convolutional Neural Network and Genetic Algorithms», *IEEE Access*, 9, 91670-91685, 2021, doi:10.1109/ACCESS.2021.3091376.
- [42] S. Sarica, J. Luo, «Stopwords in technical language processing», *Plos One*, 16(8), 1-13, 2021, doi:10.1371/journal.pone.0254937.
- [43] C.J. Harrison, C.J. Sidey-Gibbons, «Machine learning in medicine: a practical introduction to natural language processing», *BMC Medical Research Methodology*, 21(1), 1-11, 2021, doi:10.1186/s12874-021-01347-1.
- [44] P. López-Úbeda, A. Pomares-Quimbaya, M.C. Díaz-Galiano, S. Schulz, «Collecting specialty-related medical terms: Development and evaluation of a resource for Spanish», *BMC Medical Informatics and Decision Making*, 21(1), 1-17, 2021, doi:10.1186/s12911-021-01495-w.
- [45] W. Wang, A. Feng, «Self-Information Loss Compensation Learning for Machine-Generated Text Detection», *Mathematical Problems in Engineering*, 2021, 1-7, 2021, doi:10.1155/2021/6669468.
- [46] A.T. Bako, H.L. Taylor, K. Wiley, J. Zheng, H. Walter-McCabe, S.N. Kasthurirathne, J.R. Vest, «Using natural language processing to classify social work interventions», *American Journal of Managed Care*, 27(1), 1-18, 2021, doi:10.37765/AJMC.2021.88580.
- [47] V. Kumar, D.R. Recupero, D. Riboni, R. Helaoui, «Ensembling Classical Machine Learning and Deep Learning Approaches for Morbidity Identification from Clinical Notes», *IEEE Access*, 9(2021), 7107-7126, 2021, doi:10.1109/ACCESS.2020.3043221.
- [48] D. Rakesh, R.J.D. Menezes, J. De Klerk, I.R. Castleden, C.M. Hooper, G. Carneiro, M. Gilliam, «Identifying protein subcellular localisation in scientific literature using bidirectional deep recurrent neural network», *Scientific Reports*, 11(1), 1-11, 2021, doi:10.1038/s41598-020-80441-8.
- [49] B. Dreyfus, A. Chaudhary, P. Bhardwaj, V.K. Shree, «Application of natural language processing techniques to identify off-label drug usage from various online health communities», *Journal of the American Medical Informatics Association*, 28(10), 2147-2154, 2021, doi:10.1093/jamia/ocab124.
- [50] M.J. Acosta, G. Castillo-Sánchez, B. Garcia-Zapirain, I. de la Torre Díez, M. Franco-Martín, «Sentiment analysis techniques applied to raw-text data from a csq-8 questionnaire about mindfulness in times of covid-19 to improve strategy generation», *International Journal of Environmental Research and Public Health*, 18(12), 2-21, 2021, doi:10.3390/ijerph18126408.
- [51] P. Fairie, Z. Zhang, A.G. D'Souza, T. Walsh, H. Quan, M.J. Santana, «Categorising patient concerns using natural language processing techniques», *BMJ health & care informatics*, 28(1), 1-9, 2021, doi:10.1136/bmjhci-2020-100274.
- [52] T. Basu, S. Goldsworthy, G. V. Gkoutos, «A sentence classification framework to identify geometric errors in radiation therapy from relevant literature», *Information (Switzerland)*, 12(4), 1-11, 2021, doi:10.3390/info12040139.
- [53] A. Borjali, M. Magnéli, D. Shin, H. Malchau, O.K. Muratoglu, K.M. Varadarajan, «Natural language processing with deep learning for medical adverse event detection from free-text medical narratives: A case study of detecting total hip replacement dislocation», *Computers in Biology and Medicine*, 129, 1-26, 2021, doi:10.1016/j.compbiomed.2020.104140.
- [54] S.K. Prabhakar, D.-O. Won, «Medical Text Classification Using Hybrid Deep Learning Models with Multihead Attention», *Computational Intelligence and Neuroscience*, 2021, 1-16, 2021, doi:10.1155/2021/9425655.
- [55] N. Ali, A.H. Abuel-Atta, H.H. Zayed, «Enhancing the performance of cancer text classification model based on cancer hallmarks», *IAES International Journal of Artificial Intelligence*, 10(2), 316-323, 2021, doi:10.11591/ijai.v10.i2.pp316-323.
- [56] N. Khamphakdee, P. Seresangtakul, «Sentiment analysis for Thai language in hotel domain using machine learning algorithms», *Acta Informatica Pragensia*, 10(2), 155-171, 2021, doi:10.18267/j.aip.155.
- [57] S. Madichetty, M. Sridevi, «A stacked convolutional neural network for detecting the resource tweets during a disaster», *Multimedia Tools and Applications*, 80(3), 3927-3949, 2021, doi:10.1007/s11042-020-09873-8.
- [58] N. Cerkez, B. Vrdoljak, S. Skansi, «A Method for MBTI Classification Based on Impact of Class Components», *IEEE Access*, 20(2017), 1-19, 2021, doi:10.1109/ACCESS.2021.3121137.
- [59] P. Kulkarni, K.N. Cauvery, «Personally Identifiable Information (PII) Detection in the Unstructured Large Text Corpus using Natural Language Processing and Unsupervised Learning Technique», *International Journal of Advanced Computer Science and Applications*, 12(9), 508-517, 2021, doi:10.14569/IJACSA.2021.0120957.

- [60] Y. Li, P. Xu, Q. Ruan, W. Xu, «Text Adversarial Examples Generation and Defense Based on Reinforcement Learning», *Tehnicki vjesnik - Technical Gazette*, 28(4), 1306-1314, 2021, doi:10.17559/TV-20200801053744.
- [61] M. Zulqarnain, R. Ghazali, M.G. Ghouse, N.A. Husaini, A.K.Z. Alsaedi, W. Sharif, «A comparative analysis on question classification task based on deep learning approaches», *PeerJ Computer Science*, 7, 1-27, 2021, doi:10.7717/PEERJ-CS.570.
- [62] A. Moreo, A. Esuli, F. Sebastiani, «Word-class embeddings for multiclass text classification», *Data Mining and Knowledge Discovery*, 1-29, 2021, doi:10.1007/s10618-020-00735-3.
- [63] W.H. Park, N.M.F. Qureshi, D.R. Shin, «Pseudo nlp joint spam classification technique for big data cluster», *Computers, Materials and Continua*, 71(1), 517-535, 2022, doi:10.32604/cmc.2022.021421.
- [64] D.T. Tolciu, C. Săcărea, C. Matei, «Analysis of patterns and similarities in service tickets using natural language processing», *Journal of Communications Software and Systems*, 17(1), 29-35, 2021, doi:10.24138/JCOMSS.V17I1.1024.
- [65] S.K. Prabhakar, H. Rajaguru, D.O. Won, «Performance Analysis of Hybrid Deep Learning Models with Attention Mechanism Positioning and Focal Loss for Text Classification», *Scientific Programming*, 2021, 1-12, 2021, doi:10.1155/2021/2420254.
- [66] H. Ali, M.S. Khan, A. AlGhadhban, M. Alazmi, A. Alzamil, K. Al-utaibi, J. Qadir, «All Your Fake Detector Are Belong to Us: Evaluating Adversarial Robustness of Fake-news Detectors Under Black-Box Settings», *IEEE Access*, 4(2016), 1-15, 2021, doi:10.1109/ACCESS.2021.3085875.
- [67] L. Shi, Y. Zhu, Y. Zhang, Z. Su, «Fault Diagnosis of Signal Equipment on the Lanzhou-Xinjiang High-Speed Railway Using Machine Learning for Natural Language Processing», *Complexity*, 2021, 1-13, 2021, doi:10.1155/2021/9126745.
- [68] L. Burdick, J.K. Kummerfeld, R. Mihalcea, «To batch or not to batch? Comparing batching and curriculum learning strategies across tasks and datasets», *Mathematics*, 9(18), 2021, doi:10.3390/math9182234.
- [69] E. Negro-Calduch, N. Azzopardi-Muscat, R.S. Krishnamurthy, D. Novillo-Ortiz, «Technological progress in electronic health record system optimization: Systematic review of systematic literature reviews», *International Journal of Medical Informatics*, 125, 1-8, 2021, doi:10.1016/j.ijmedinf.2021.104507.

APPENDIX A. NLP TECHNIQUES

Dimension	Definition
Word segmentation (Tokenize)	It is the process of converting paragraphs into inputs for the computer through a word list.
Data cleanup (Stop word)	It is the process of removing words that do not add exclusionary meaning to a sentence.
Lexicographic analysis with stemming	It is the process of converting each word of the sentence to its root form by removing or replacing suffixes.
Lexicographic analysis with lemmatization	It is a more accurate process than stemming and involves making an analysis of the vocabulary and its morphology to return to the basic form of the word.

APPENDIX B. NLP ALGORITHMS

Type	Algorithm	Definition
Basic mathematical functions	POS	It is the process of grammatical tagging or disambiguation of word categories.
	Name entity recognition (NER)	It is the process of "finding out" if a piece of data belongs to a person or business organization.
	N-gram	It is a sub-sequence of n items of a text data sequence. It is a probabilistic algorithm that allows making a statistical prediction of the next item of a sequence of a string of text data.
	Bag of words (BoW)	It is the process that allows the feature extraction from the text, determines the number of times that there is a word in the sentence.
Basic statistical algorithms	Term frequency - inverse document frequency (TF-IDF)	It is a statistical model that allows scoring the data to reflect their relevance in a given document.
	Text vectorization	It is the process which transforms the input of language into something that the computer can understand
	Statistical standardization	It is the process used to scale features of document data.

APPENDIX C. PLN ALGORITHMS OF THE 46 ARTICLES.

Author	PLN Algorithms	ML Algorithms
[17]	BoW, TF-IDF, N-Gram	SVM, NB, K-Medias
[63]	TF-IDF	SVD, R
[43]	TF-IDF	R, SVM, ANN
[44]	N-gram, TF-IDF	RF, KNN, AD
[48]	CBoW	LSTM, RNN
[9]	TF-IDF	R, LSTM, NB, RF, SVM
[30]	TF-IDF	NB, RF, AD
[16]	BoW, TF-IDF	ANN, RNN, LSTM
[49]	TF-IDF	SVM, NB, RF, RNN
[68]	Word2Vect	LSTM

[32]	CBoW, TF-IDF	CNN, LSTM
[42]	TF-IDF	LSTM
[33]	Glove	LSTM
[60]	Word2Vect	NB, SVM, MLP, CNN, RNN, LSTM, GRU
[34]	N-Gram	R, NB, RF
[50]	TF-IDF, Glove, Skip Gram	MLP
[51]	BoW	LDA
[52]	BoW, Skip Gram	SVM
[28]	Word2vect, Glove	RNN, LSTM
[29]	N gram	MLP
[64]	BoW, Word2vec, Glove	LSTM
[24]	TF-IDF, Word2Vec	SVM
[53]	N gram	KNN, RF, SVM, CNN, LSTM
[35]	BoW	NB
[36]	TF-IDF	NB, MLP, SVM, KNN
[59]	BoW, Word2Vect	LSTM
[65]	Word2Vect	RNN, LSTM
[58]	TF-IDF, BoW, Glove	NB, KNN, SVM, LSTM, R
[41]	Glove	CNN
[37]	TF-IDF, Word2Vec	LSTM, CNN
[56]	Word2Vec, BoW, TF-IDF	SVM, NB, R, RF, AD, PA, ADA
[54]	Word2Vec, Glove	LSTM, GRU
[38]	Glove	KNN, NB, PA, LSTM
[61]	Word2Vec	LSTM, GRU, CNN
[67]	BoW	LDA, SVM, NB, R, RF, KNN
[39]	TF-IDF	NB, R, AD, RF
[40]	Glove	LSTM
[66]	Glove	MLP, CNN, RNN
[25]	TF-IDF, N Gram, BoW	LDA, GRU, CNN
[55]	Glove	SVM, RNN, CNN
[45]	Word2Vec	CNN, RNN, LSTM
[62]	TF-IDF, Glove	SVM, CNN, LSTM
[46]	TF-IDF, Glove	NB, R, SVM, LSTM
[47]	TF-IDF, Word2Vec, Glove	SVM, KNN, NB, RF, AD, LSTM
[27]	TF-IDF, Word2Vec	LSTM
[57]	Word2Vec	SVM, KNN, CNN

APPENDIX D. MACHINE LEARNING ALGORITHMS – ML

ML Type	Algorithm	Overview
Supervised learning	K-NN	It is an algorithm that can be used to classify new samples or to predict values by looking for the “most similar” data points (by proximity).
	Regression - R	It is the algorithm that determines the relationships between dependent and independent variables for prediction and prognosis.
	Decision tree - DT	It is an algorithm that uses the fork for every possible outcome of a decision.
	Support Vector Machines - SVM	It is an algorithm that seeks to find a hyperplane that best separates two different kinds of data points.

	Naive Bayes - NB	It is a classification algorithm based on Bayes' theorem and classifies each value as independent from any other. It uses probability to predict a class or category.
	Radom Forest - RF	It is the algorithm that represents a set of decision trees in which each tree trains with different data samples from the same problem.
	Passive aggressive – PA	It is an algorithm that is used for large-scale learning. Input data come in sequential order and the machine learning model is updated step by step.
	Singular Value Decomposition – SVD	It is the algorithm used to eliminate redundant data. It determines which values are important and removes those that are not.
	AdaBoost – ADA	It is an algorithm that can be used together with other learning algorithms to improve its performance.
Unsupervised learning	Cluster K-Media	It is an algorithm that trains and properly knows the data to find hidden groups.
	Latent Dirichlet Allocation – LDA	It is an algorithm and its objective is to find the topics to which a document belongs based on the words it contains.
Deep Learning	Convolutional neural networks -CNN	It is one of the variants of neural networks that is widely used in the computer vision field.
	Recurrent Neural Networks – RNN	It is an algorithm widely used in natural language processing. It is used to analyze time series data.
	Long Short Term Memory - LSTM	It is an algorithm that introduces loops in the network diagram to memorize previous states of variables to decide which one will be next.
	Artificial neural network - ANN	It is a group of multiple perceptrons/neurons in each layer.
	Multilayer Perceptron - MLP	It is a network class consisting of at least three layers of nodes: an input layer, a hidden layer, and an output layer.
	Gating Circulation Unit – GRU	It is an enhanced RNN.

Hybrid Fault Diagnosis Method based on Wavelet Packet Energy Spectrum and SSA-SVM

Jinglei Qu¹, Xiaojie Ma³, Mengmeng Wang⁴

School of Mechanical Engineering
Henan Institute of Technology, Xinxiang, China

Bingxin Ma²

School of Materials Science and Engineering
Henan Institute of Technology, Xinxiang, China

Abstract—As one of the important components of mechanical equipment, rolling bearing has been widely used, and its motion state affects the safety and performance of equipment. To enhance the fault feature information in the bearing signal and improve the classification accuracy of support vector machine, a hybrid fault diagnosis method based on wavelet packet energy spectrum and SSA-SVM is proposed. Firstly, the wavelet packet decomposition is used to decompose vibration signals to generate frequency band energy spectrum, and the bearing characteristic information is constructed from the energy spectrum to extract and enhance the bearing fault characteristic information. Secondly, the penalty and kernel parameters are optimized globally by sparrow search algorithm to improve the classification accuracy of support vector machine, and then construct the WPES-SSA-SVM model. Finally, the proposed model is used to diagnose and analyze the measured signals. Compared with BP, ELM and SVM, the effectiveness and superiority of the proposed method are verified.

Keywords—Wavelet packet energy spectrum; sparrow search optimization; support vector machine; rolling bearing

I. INTRODUCTION

With the deep integration of new generation information technology and manufacturing industry, the mechanical equipment is becoming more and more complex, accurate and intelligent. With the continuous operation of mechanical equipment, its running state and key parts will gradually degenerate, and the probability of failure and shutdown will gradually increase, which will affect the normal production and processing of enterprises. As one of the important components of machinery, rolling bearings are widely used because of their convenient use and maintenance, reliable operation and good starting performance [1]. Using the characteristics of bearings, the sliding friction between parts is transformed into rolling friction, which improves the production efficiency of the equipment. Once damaged, it will lead to problems in the operation of mechanical equipment, reduce the working efficiency, and even cause the functional failure of rotating machinery, resulting in serious economic losses and personal casualties[2-3]. Therefore, it is of great practical value to timely find and take corresponding measures for the faults of rolling bearings, and it has become a research hotspots in intelligent fault diagnosis.

In recent years, fault diagnosis methods for rolling bearings have been emerging and developing[4-7]. Fault diagnosis methods for rolling bearings have mushroomed and developed continuously. In general, the fault diagnosis

techniques of rolling bearings include: based on vibration signal[8], acoustic signal[9], electrical signal[10] and temperature signal[11]. Among them, vibration signal is more widely used, more intuitive and simple, because it can best represent the fault characteristic information in the process of bearing operation.

As the rapid and continuous development of machine learning and artificial intelligence, more and more researchers combine bearing fault diagnosis with it, and the intelligent fault diagnosis methods and systems are gradually improved. The common fault identification methods include deep learning (DL)[12], artificial neural network (ANN)[13], decision tree (DT)[14] and support vector machines (SVM)[15,16]. Literature [17] proposed the improved BP neural network algorithm Levenberg-Maquardt algorithm in order to improve the diagnostic efficiency of BP neural network. Literature [18] proposed a fault extraction method based on modified Fourier mode decomposition (MFMD) and multi-scale displacement entropy, and combined with BP neural network. Experiments show that this method has high recognition accuracy for different types of faults. In literature [19], wavelet packet energy and decision tree algorithm are combined to extract faults using wavelet packet energy, and then faults are identified and classified using decision tree model. In view of the low fault diagnosis rate of rolling bearings, the method of wavelet packet decomposition and gradient lifting decision tree (GBDT) was proposed in literature [20], and the extracted fault feature data set was input into the classification model of gradient lifting decision tree for fault diagnosis. In literature [21], scale invariant feature transform (SIFT) and kernel principal component analysis (KPCA) were used to extract faults, and SVM classifier was combined to achieve fault classification. Literature [22] applied SVM to fault state identification of rolling bearings and achieved good results. Literature [23] proposed a rolling bearing fault diagnosis method optimized by simplex evolutionary algorithm and SVM. Literature [24] diagnoses fault types by reducing high-dimensional data and using LSSVM.

At present, various intelligent optimization algorithms have emerged one after another, such as particle swarm optimization (PSO), whale optimization algorithm (WOA), ant colony optimization (ACO), genetic algorithm (GA), sparrow search algorithm (SSA), etc., and the combination and improvement of other algorithms have also achieved good results[25]. In reference [26], PSO was used to optimize SVM to realize the identification of multiple fault states of rolling

bearings. In [27], gray wolf optimization algorithm (GWO) was used to optimize the kernel function parameters of SVM globally, so as to achieve the best classification performance of SVM and improve the accuracy of classification recognition. Aiming at the influence of mixed noise of bearing vibration signals on useful information extraction, an optimization classifier based on multi-scale permutation entropy and cuckoo search algorithm (CS) was proposed in literature [28], which used CS to optimize the global optimal solution of SVM. Literature [29] proposed a method based on quantum behavior particle swarm optimization algorithm (QPSO), multi-scale displacement entropy and SVM to construct fault feature sets to realize fault identification of rolling bearings. Compared with single method for fault diagnosis, the combinatorial optimization methods have higher accuracy, but at the same time, different optimization methods have different problems, for example, BP model must be learned through a large amount of sample data, even if it has been optimized the BP network parameters globally by optimization algorithm, the model is still not ideal in a small sample environment. SVM parameters can be optimized by PSO and other optimization algorithms to improve the classification accuracy, but this algorithm is prone to fall into local extremum. Therefore, combining the advantages of each algorithm and joint application to improve the effectiveness of rolling bearing status identification and fault diagnosis is the current research trend.

To improve the accuracy of bearing fault diagnosis, this paper firstly uses wavelet packet energy spectrum to extract energy spectrum feature vectors of bearing vibration signals, which are used as the input of SVM. Meanwhile, SSA algorithm is used to optimize the parameters of SVM globally, so as to build a hybrid model. The feasibility and effectiveness of the model are verified by experiments.

The rest parts of this paper are given as lists: Section 2 presents the preliminaries. Section 3 describes of the proposed method. Section 4 details the experimental setup. Section 5 analyzes and discusses the experimental results. Finally, Section 6 outlines the main conclusions.

II. PRELIMINARIES

A. Wavelet Packet Energy Spectrum

Wavelet packet decomposition can decompose signals into different frequency bands without leakage and overlap according to any time-frequency resolution. After wavelet packet transform, the information is intact and all frequencies are retained, which provides strong conditions for extracting the main information in the signal. This decomposition can be performed as many times as needed to obtain the desired frequency.

Fig. 1 shows the schematic diagram of orthogonal wavelet packet decomposition of a signal. The original signal was denoted as S_0 , and the two sub-bands S_{10} and S_{11} of layer 1 can be obtained after wavelet packet decomposition through filters H and G. Decompose the two sub-components of the first layer respectively to obtain the four sub-bands S_{20} , S_{21} , S_{22} and S_{23} of the second layer; By analogy, the sub-band of layer n can finally be obtained.

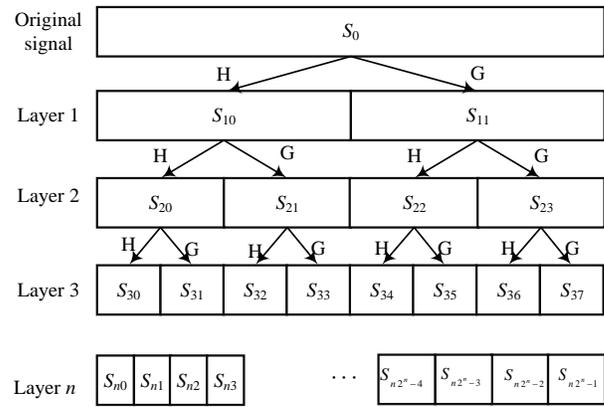


Fig. 1. Schematic Diagram of Wavelet Packet Decomposition.

As can be seen from Fig.1, wavelet packet decomposition decomposes the decomposed frequency band several times, and re-decomposes the high frequency part without subdivision in the wavelet decomposition. In addition, according to the characteristics of the signal to be decomposed, the corresponding sub-frequency band can be adaptively selected to match the frequency spectrum of the signal. After wavelet decomposition, all the characteristic information, including the low frequency part and the high frequency part, can be preserved, which provides strong support for the feature information extraction of the signal.

It can also be seen from Fig.1 that if there are too many decomposition layers, the dimension of the data to be processed will be increased and the unrestricted decomposition cannot continue. In practical application, it is necessary to select an appropriate decomposition level according to the actual situation.

Wavelet packet energy spectrum enhances the stability of wavelet packet decomposition coefficient by extracting the energy of sub-band to construct feature vector. The wavelet packet frequency band energy is defined as follows:

Using wavelet packet to decompose the original signal S_0 in n level, and $2n$ sub-frequency band can be decomposed. The energy calculation formula of sub-frequency band ni is Formula 1.

$$E(ni) = \sum f_{ni}^2 \tag{1}$$

where, f_{ni} is the coefficient of sub-frequency band ni , $ni = 1, 2, \dots, 2^n - 1$.

Therefore, the wavelet packet frequency band energy spectrum is defined as Formula 2.

$$En = [En_0, En_1, En_2, \dots, En_{2^n-1}] \tag{2}$$

B. Support Vector Machines

SVM is a machine learning algorithm based on statistical learning theory, which can successfully deal with many data mining problems such as pattern recognition, classification and regression analysis. It shows many unique advantages in solving small sample, nonlinear and high-dimensional pattern recognition problems, and overcomes the problems of dimension disaster and over-learning to a large extent.

Based on the theory of minimum construction risk, support vector machine maximizes the distance between the elements closest to the hyperplane and the hyperplane. Its core is to establish the best classification hyperplane, so as to improve the generalization processing ability of learning classification machine.

Taking binary classification as an example, its basic idea can be summarized as follows: first map the input vector to a high-dimensional feature space through some prior selected nonlinear mapping such as kernel function, and then seek the optimal classification hyperplane in the feature space, enables it to as much as possible to separate two classes of data points correctly, at the same time to separate two classes of data point furthest distance classification surface, as shown in Fig. 2.

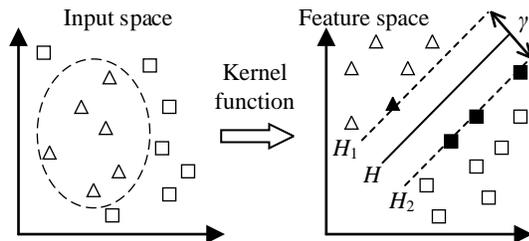


Fig. 2. Classification Principle of SVM Method.

In Fig. 2, square and triangle represent two types of samples respectively. H is the optimal classification hyperplane; H_1 and H_2 are straight lines that pass through the boundary points of the two types of samples and are parallel to H , and the distance between them γ is the interval. The optimal classification line requires that the classification line can not only correctly classify the two categories, but also maximize the interval. The vector closest to the optimal classification hyperplane is called the support vector.

Assume the training sample set $\{x_i, y_i\}, i = 1, 2, \dots, m$; $x_i \in R^n, y_i \in \{-1, +1\}$, where x_i is the input index, y_i is the output index, m is the sample number, and n is the characteristic dimension of the sample. In the case of linear divisibility, there is a hyperplane that separates the two types of samples completely, as shown in Formula 3.

$$(\omega \cdot x) + b = 0 \quad (3)$$

where, $\omega = (\omega_1, \omega_2, \dots, \omega_n)$ is the weight vector of the training sample, which determines the direction of the hyperplane. x is the input vector; b is the distance between the hyperplane and the origin.

Solving the optimal classified hyperplane is to find the optimal ω and b , therefore, it can be summed up as the following quadratic programming problem:

$$\begin{cases} \min \frac{1}{2} \|\omega\|^2 \\ y_i = [(\omega \cdot x_i) + b]; i = 1, 2, \dots, m \end{cases} \quad (4)$$

In order to solve the quadratic programming problem of Formula 4, the Lagrange function \mathbf{a} is introduced and the duality principle is used to transform the original optimization problem into Formula 5:

$$\begin{cases} \max Q(\mathbf{a}) = \sum_{i=1}^m a_i - \frac{1}{2} \sum_{i=0}^m \sum_{j=0}^m a_i a_j y_i y_j x_i^T x_j \\ s. t. \sum_{i=1}^m a_i y_i = 0; a_i \geq 0; i = 1, 2, \dots, m \end{cases} \quad (5)$$

According to Formula 5, the optimal \mathbf{V} is $\mathbf{a}^* = (a_1^*, a_2^*, \dots, a_m^*)^T$. Thus, the optimal weight vector ω^* and the optimal value b can be calculated by Formula 6 and Formula 7.

$$\omega^* = \sum_{i=1}^m a_i^* y_i x_i \quad (6)$$

$$b^* = y_i - \sum_{j=1}^m a_j^* y_j x_j^T x_i \quad (7)$$

Then the optimal classification hyperplane is $(\omega^* \cdot x) + b^* = 0$, and the optimal classification function is obtained.

$$f(x) = \text{sgn}(\sum_{i=1}^m a_i^* y_i (x_i \cdot x_j) + b^*), x \in R^n \quad (8)$$

C. Sparrow Search Algorithm

SSA realizes optimization based on the idea that swarm organisms in nature can obtain a better living environment through mutual cooperation[30]. The bionic principle is as follows: in order to obtain abundant food, the sparrow population is divided into explorers and followers in the process of foraging. The explorer in the sparrow population who finds abundant food sources is responsible for providing the foraging area and the direction of food sources for the population, and the followers is responsible for finding more food according to the location provided by the explorer. At the same time, individual sparrows will also monitor the behavior of other individuals and compete for supplies with high-foraging peers. When the population is in danger, it will make anti predation behavior. The external sparrow will constantly adjust its position to move closer to an internal or adjacent partner to increase its own security. Therefore, the distribution of food in space can be regarded as the numerical value of function in three-dimensional space. The purpose of sparrow search is to find the global optimal value.

The specific implementation process of sparrow search algorithm is as follows. In the process of searching for food, the randomly generated position matrix X of n sparrows in the d dimensional space is shown as follows:

$$X = \begin{bmatrix} x_1^1 & x_1^2 & \dots & x_1^d \\ x_2^1 & x_2^2 & \dots & x_2^d \\ \vdots & \vdots & \ddots & \vdots \\ x_n^1 & x_n^2 & \dots & x_n^d \end{bmatrix} \quad (9)$$

where n represents the number of sparrows, d represents the dimension of the variable of the problem to be optimized, x_i^j ($i = 1, 2, \dots, n; j = 1, 2, \dots, d$) is the position of the j sparrow in i -dimensional space.

The fitness values are calculated and sorted to determine the finders and entrants, and 10% of randomly selected individuals are scouters. Obtain the current optimal sparrow individual position, and the best fitness value. For the first generation of sparrows, the initial optimal is obtained.

$$F = \begin{bmatrix} f([x_1^1 & x_1^2 & \dots & x_1^d]) \\ \vdots \\ f([x_n^1 & x_n^2 & \dots & x_n^d]) \end{bmatrix} \quad (10)$$

where f represents fitness values of individual sparrows.

In constant iterative optimization process, the explorers in the sparrow population have two main tasks: looking for food and guiding the movement of the population. When the scouters feel dangerous, will alert the populations and guide the followers to a safe area. The location of the explorers is updated as follows:

$$X_{ij}^{t+1} = \begin{cases} X_{ij}^t \cdot \exp\left(\frac{-i}{\alpha \cdot T}\right), & \text{if } R_2 < ST \\ X_{ij}^t + QL, & \text{if } R_2 \geq ST \end{cases} \quad (11)$$

where X_{ij}^t represents the position of the i -th sparrow in the j -th dimension of the t generation. α is a random number in the range of $[0,1]$; T represents the maximum number of iterations; Q is a random number that follows normal distribution; L represents a $1 \times d$ matrix where each element is 1; R_2 and ST represents the alarm value and alarm threshold respectively, $R_2 \in [0,1]$, $ST \in [0.5,1]$. When $R_2 < ST$ means that there are no predators around foraging at this time and the explorer can conduct extensive foraging operation. Conversely, it indicates that some sparrows in the group have found predators and send Danger Warnings to the rest, thus ensure that all sparrows can quickly move to a safe area to forage.

Followers search for food by monitoring and following the explorers with the highest fitness. According to the sorting principle, when $i > n/2$, the individual fitness value is low, and these followers need to search other locations to improve the individual fitness value. Conversely, the sparrow will randomly find a location near the current optimal location for feeding.

$$X_{ij}^{t+1} = \begin{cases} Q \cdot \exp\left(\frac{X_{worst}^t - X_{ij}^t}{i^2}\right), & \text{if } i > n/2 \\ X_p^{t+1} + |X_{ij}^t - X_p^{t+1}| \cdot A^+ \cdot L, & \text{if } i \leq n/2 \end{cases} \quad (12)$$

$$A^+ = A^T(AA^T)^{-1} \quad (13)$$

where X_{worst}^t represents the global worst position of the t -th iteration; X_p^{t+1} is the best position of the $t+1$ generation explorer. A is a $1 \times d$ dimensional matrix with each dimensional value randomly generated from 1 or -1.

Individual sparrows will move to the search circle or other companions when they encounter danger during the foraging process. The method of updating the position of individual sparrows in this process is shown in Formula (14).

$$X_{ij}^{t+1} = \begin{cases} X_{best}^t + \beta \cdot |X_{ij}^t - X_{best}^t|, & \text{if } f_i > f_g \\ X_{ij}^t + K \cdot \left(\frac{|X_{ij}^t - X_{worst}^t|}{(f_i - f_w) + \varepsilon}\right), & \text{if } f_i = f_g \end{cases} \quad (14)$$

where β is the step size control parameter, and it follows the normal distribution with mean value 0 and variance 1; K is the moving direction of the sparrow, and the value range is $[-1, 1]$; ε is the minimum constant to avoid zero denominator; X_{best}^t represents the current global optimal location; f_i represents the fitness value of sparrow i ; f_w and f_g represent the current worst and best fitness values, respectively.

III. PROPOSED MODEL

To improve the fault diagnosis accuracy of bearing vibration signals, a hybrid fault diagnosis model is constructed by using wavelet packet energy spectrum, SSA and SVM, which is named WPES-SSA-SVM. In order to accurately extract features, wavelet packet energy spectrum is used to extract feature information from vibration signals, and the energy of reconstructed signals are calculated through wavelet packet decomposition and reconstruction, and the feature vector is established. Then, SSA is used to optimize the penalty parameter c and kernel parameter g globally to improve the learning ability and generalization ability of SVM classifier. The model consists of data feature extraction, SSA optimization and SVM recognition. The functions of each part and the information transmission between them are shown in Fig. 3.

1) *Data feature extraction module*: Using wavelet packet decompose the bearing vibration signal, and the wavelet packet frequency band energy spectrum is generated according to the decomposition results. Taking the energy spectrum information as the fault diagnosis features, and divide it into training and test data set in proportion. Then, the training data is transmitted to the SSA optimization module, and the training and test data are transmitted to the SVM recognition module.

2) *SSA optimization module*: The SSA optimization module receives the training data from the data feature extraction module and the value range of penalty parameter c and kernel parameter g from the SVM recognition module respectively, uses SSA to find the best penalty parameter c and kernel parameter g , and returns them to the SVM recognition module.

3) *SVM recognition module*: The SVM recognition module first transmits the value range of penalty parameter c and kernel parameter g to the SSA optimization module for parameter optimization, then receives the optimized parameters, and carries out machine training using the training data received from the data feature extraction module. After that, the fault diagnosis on the test data is recognized to test the recognition effect.

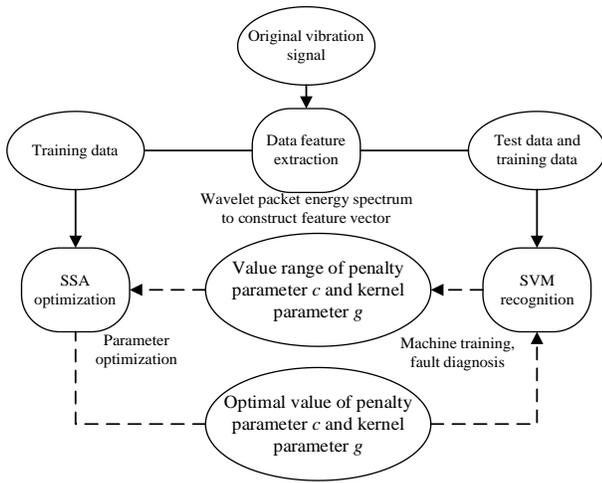


Fig. 3. Function and Information Transmission Path of each Module in WPES-SSA-SVM Model.

The algorithm of the model is divided into nine steps, and the flow chart is shown in Fig. 4.

Step 1: The original vibration signal is decomposed by wavelet packet, and the frequency band energy spectrum is calculated, and then the data is randomly divided into test data and training data in proportion.

Step 2: Select the kernel function to construct SVM, mainly including linear kernel function, RBF kernel function, polynomial kernel function and Sigmod kernel function, and set the value range of penalty parameter c and kernel parameter g .

Step 3: Initialize sparrow population. Set the population size $Size$, the maximum number of iterations T_{max} , the individual position X , where X is the multidimensional coordinate composed of penalty parameter c and kernel parameter g , the proportion E, F, S of explorers, followers and scouts, and the safety threshold ST .

Step 4: Using the classification accuracy as the fitness function value of SSA.

Step 5: Find the global optimal position. The fitness value f of individual position is obtained by using training data. The larger the value is, the better the position is, and the global optimal position is the position with the largest f . If multiple positions at the same f , the optimal position is the one with the smallest penalty parameter c .

Step 6: Update the population position and global optimal position.

Step 7: Iteration number condition judgment. If the current of iterations $t < T_{max}$, return Step 6 to continue running; Otherwise, execute Step 8.

Step 8: Using SSA optimization to get the best parameters, and the SVM is trained through the training data.

Step 9: Input the test data into SVM, output the calculated bearing fault label value, identify the fault type, and compare it with the real fault type label in the original data to verify the diagnosis effect.

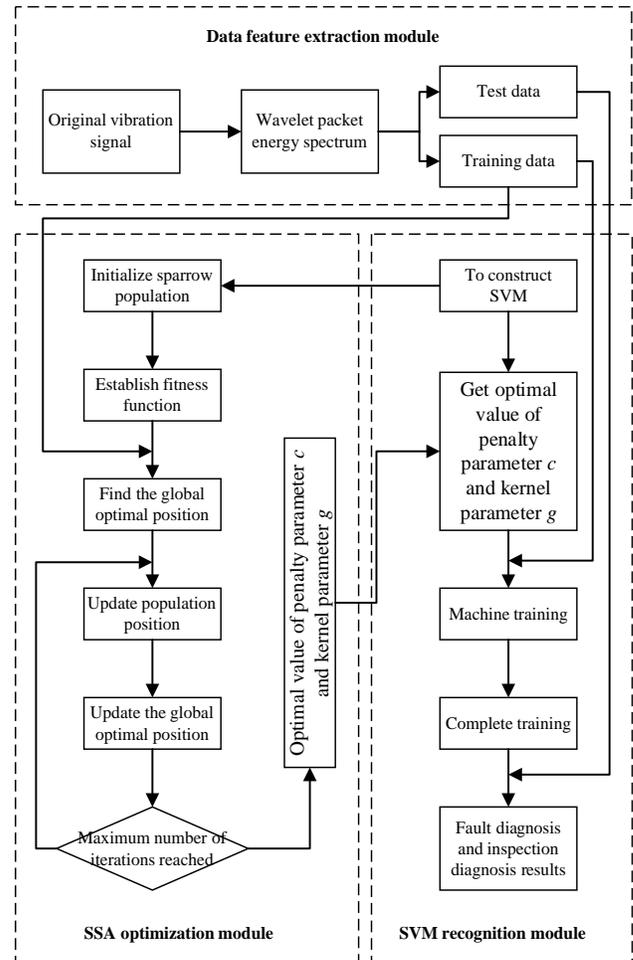


Fig. 4. Algorithm Flow Chart of WPES-SSA-SVM Model.

IV. EXPERIMENTATION

Feature extraction and fault diagnosis were performed using simulated fault data from the bearing experiment data provided by Case Western Reserve University (CWRU). The data set has been applied in many experimental studies and achieved good results. The time domain and wavelet packet characteristics of vibration signals are extracted from the official experimental data and fault diagnosis is carried out. The structure of the bearing test bench is shown in Fig. 5 [31].

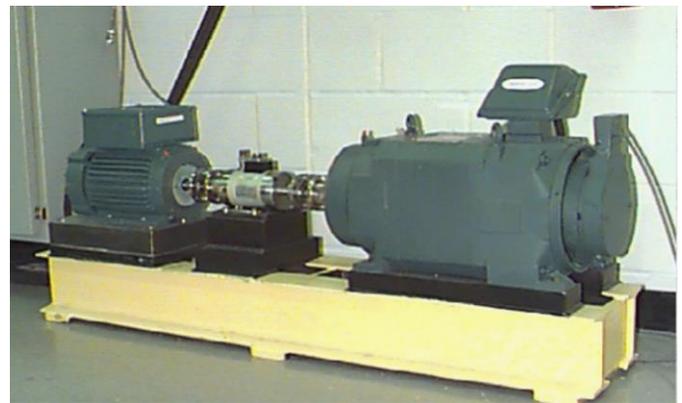


Fig. 5. CWRU Bearing Test Bench.

The test bench is composed of three-phase induction motor, torque sensing device, electronic control unit, dynamometer and intermediate shaft. During the experiment, the motion state of the rolling bearing in the actual work is simulated. Single point defects with different widths such as 0.007, 0.014, 0.021, 0.028 and 0.040 inch are machined on different parts of the bearing by spark machining technology, so as to obtain the experimental data of different fault types, such as rolling element, inner race and outer race fault.

In this paper, the fault body diameter is 0.007 inch, the motor load horsepower is 1hp, the bearing model is SKF-6205-2RS-JEM, and the sampling frequency of acceleration sensor is 48KHz to collect the vibration signal data of normal bearing, inner race fault, outer race fault and rolling element fault at the driving end. Take 100 groups of data samples for each state, with a total of 400 groups of data samples. 100 samples are randomly divided into 70 training samples and 30 test samples after feature extraction by wavelet packet energy spectrum. The training samples are used to extract features for classification model training, and the test samples are used to test the effect of classification model. The parameters of rolling bearing are shown in Table I. The division and label setting of experimental data are shown in Table II.

TABLE I. ROLLING BEARING PARAMETERS

Type	Parameter
Bearing model	SKF-6205-2RS-JEM
Inner diameter	25.00mm
Outer diameter	52.00mm
Rolling elements number	9
Rolling element diameter	7.94mm
Pitch diameter	39.04mm
Contact angle	90°

TABLE II. STATISTICAL TABLE OF EXPERIMENTAL DATA

Bearing state	Fault body diameter	Number of samples	Label
Normal	0.007	100	1
Inner race fault	0.007	100	2
Outer race fault	0.007	100	3
Rolling element fault	0.007	100	4

V. RESULT AND DISCUSSION

The time domain waveform diagram can intuitively observe the waveform distribution and amplitude of the vibration signal in each state. The waveform will fluctuate with the fault location and size. The vibration signals of normal and different faults of bearings are shown in Fig. 6.

The wavelet packet decomposition with wavelet basis function as db3 is used to decompose the normal state, inner race, outer race and rolling element fault signals respectively, so as to obtain the decomposition coefficient and reconstruction coefficient, and then use the reconstruction

coefficient to reconstruct, finally obtain 8 sub-band energy, and the energy proportion of each frequency band is analyzed. Due to space limitation, this paper only lists the wavelet packet components of reconstructed nodes in normal state, as shown in Fig. 7. The energy proportion of 8 sub-bands in different states is shown in Fig. 8.

It can be clearly seen from Fig. 8 that there are differences in normalized amplitude of wavelet energy spectrum in different frequency bands after reconstruction of each node. Among them, the energy spectrum of sub-band 1 and 2 is relatively large in the four states, followed by the energy spectrum of sub-band 3 and 4, and the energy spectrum of sub-band 5, 6, 7 and 8 is relatively small, but there are slightly different in different states.

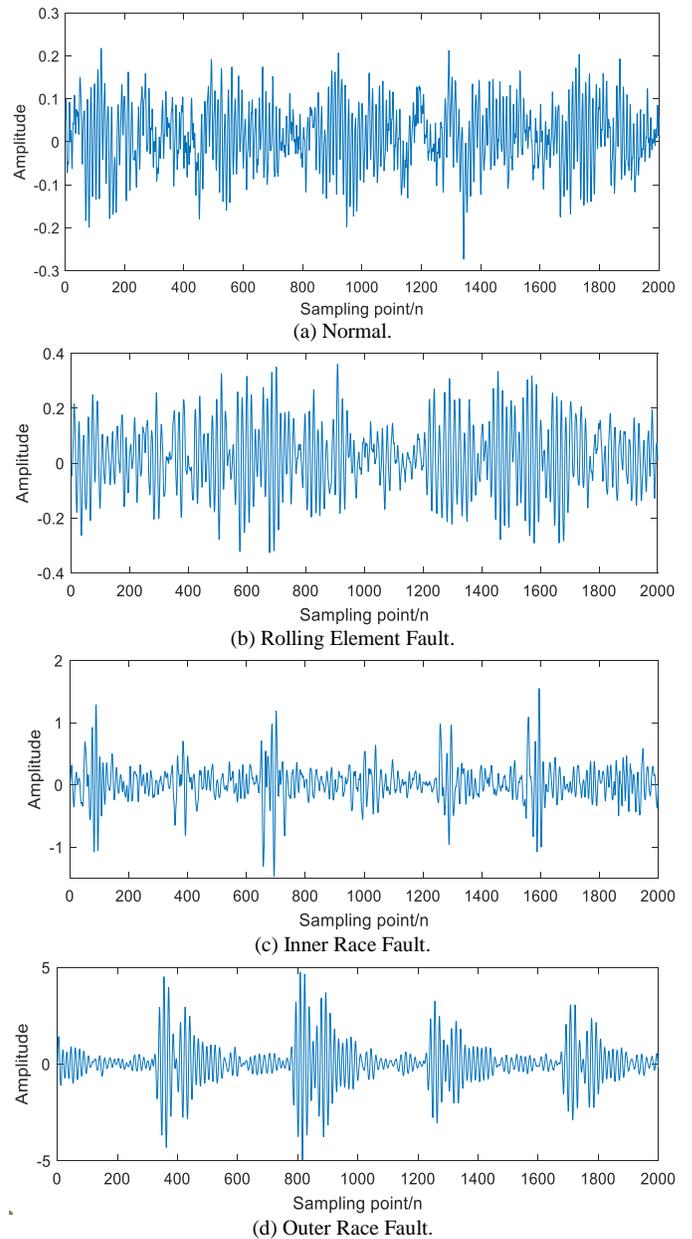


Fig. 6. Bearing Vibration Signal Diagram.

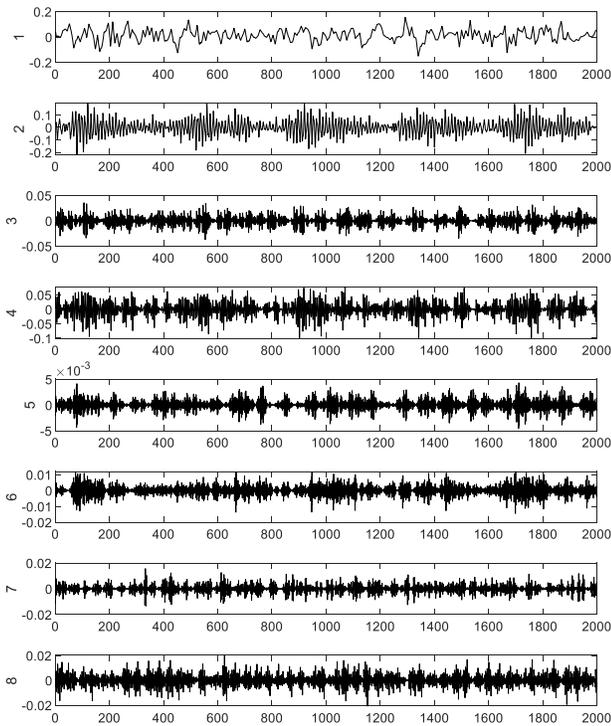
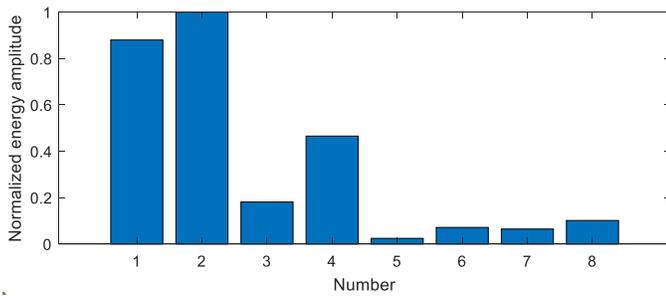
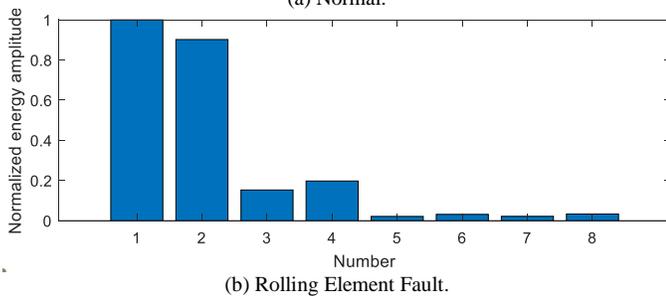


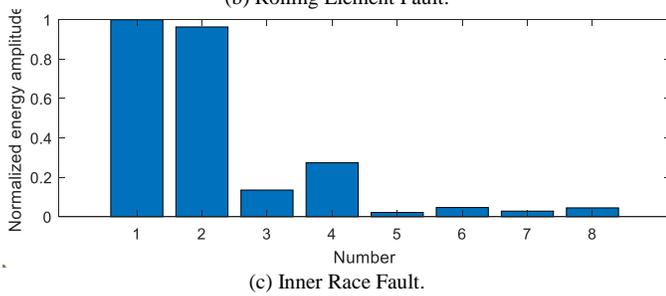
Fig. 7. Wavelet Packet Component of Normal Bearing Vibration Signal.



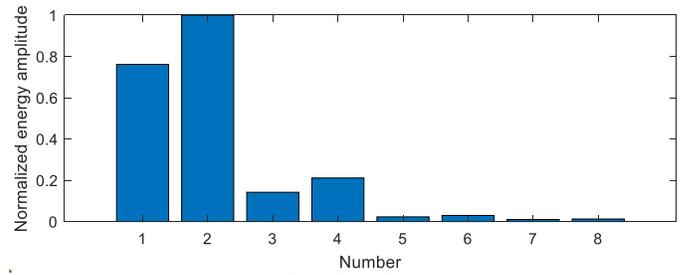
(a) Normal.



(b) Rolling Element Fault.



(c) Inner Race Fault.



(d) Outer Race Fault.

Fig. 8. Energy Spectrum of Wavelet Packet Frequency Band.

For example, when the bearing is in the normal state, the energy spectrum value of sub-band 4 is higher than that in the fault state. When the outer race fault occurs, the energy spectrum value of sub-band 1 is lower than that in other cases. In case of bearing inner race fault or rolling element fault, the energy spectrum graph is relatively close, but there is still a certain gap between the values of sub-band 4 and sub-band 6. The difference of wavelet packet energy spectrum graphics in different states reflects that the features extracted by wavelet packet transform are sensitive to the fault feature information of vibration signal. Therefore, the energy amplitude corresponding to each sub-band and the energy difference between frequency bands can be used to evaluate the different states of bearings.

To verify the feasibility and effectiveness of WPES-SSA-SVM, experiments were conducted on BP, ELM, SVM and WPES-SSA-SVM respectively. The diagnosis results are shown in Fig. 9, where 'o' stands for the fault category of the actual testing set, and '*' stands for the fault category predicted by the model.

From Fig. 9, the BP model misjudged 18 faults in total, including 5 rolling element faults misjudged into 3 inner race faults and 2 outer race faults, 8 inner race faults misjudged into 3 outer race faults and 4 rolling element faults and 1 normal, 5 outer race faults misjudged into 2 inner race faults, 2 rolling element faults and 1 normal, and the diagnostic accuracy is 85%. The ELM model misjudged a total of 16 faults, of which 4 rolling element faults were misjudged as 1 inner race fault and 3 outer race faults, 6 inner race faults were misjudged as 3 rolling element faults and 3 outer race faults, 6 outer race faults were misjudged as 1 rolling element fault and 5 inner race faults, and the diagnostic accuracy was 86.67%. There are 14 wrong judgments in SVM model, including 3 wrong judgments of rolling element fault, 2 wrong judgments of inner race and 9 wrong judgments of outer race. The diagnostic accuracy is 88.33%. WPES-SSA-SVM model misjudged 4 faults in total, including 1 rolling element fault misjudged as inner race fault, 2 inner race faults misjudged as outer race fault and 1 outer race fault misjudged as rolling element fault. The number of misjudged in the four states has been well improved. WPES-SSA-SVM model has the best diagnostic effect for ELM model, SVM model and BP model, and the diagnostic accuracy is 96.67%. The experimental results show that using wavelet packet energy spectrum for feature extraction and SSA to optimize SVM model can improve the performance of fault diagnosis, and has obvious advantages over other non-optimized models.

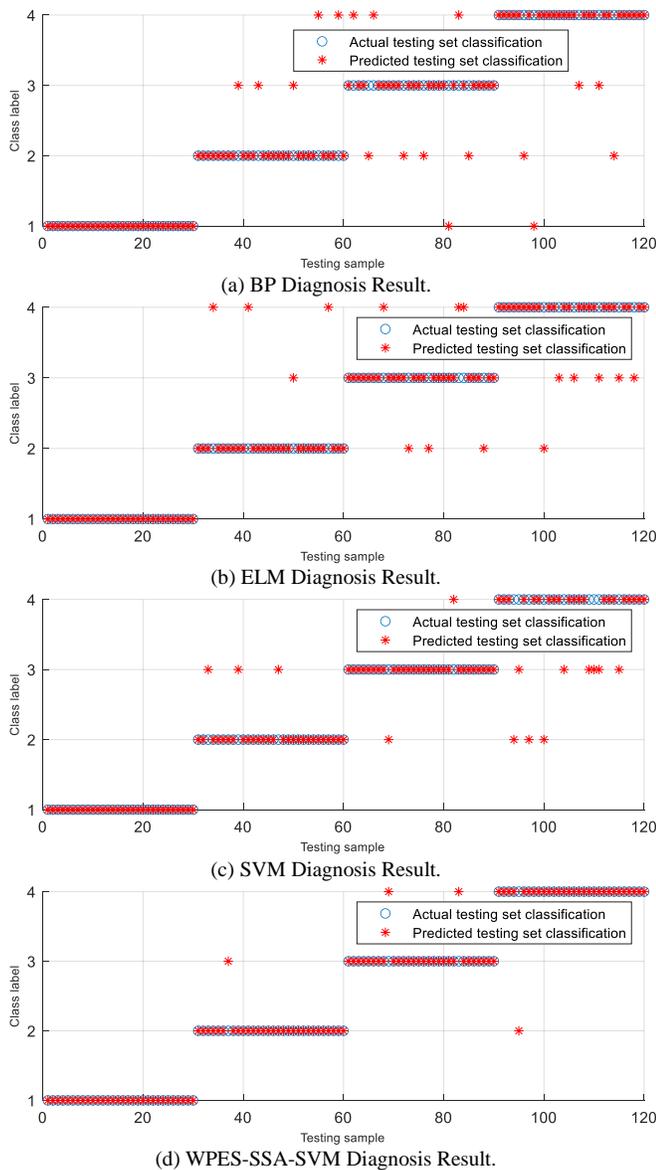


Fig. 9. Fault Diagnosis Results of Different Models.

VI. CONCLUSION

In this paper, we proposed a hybrid fault diagnosis method based on wavelet packet energy spectrum, SSA, and SVM in rolling bearing. Aiming at the difficulty of feature extraction of bearing vibration signals, wavelet packet decomposition was used to extract the wavelet packet features of vibration signals, and the energy spectrum of wavelet components is calculated and normalized to form the feature vector set, which fully contained the fault feature information of vibration signals. To improve the accuracy of fault diagnosis, the penalty parameter c and kernel parameter g of SVM are optimized by using the good global optimization ability of SSA, so as to build a hybrid fault diagnosis model WPES-SSA-SVM. To verify the classification performance of WPES-SSA-SVM, the CWRU bearing vibration data set is used to extract fault features and diagnose faults. The results show that compared with BP, ELM, and SVM, the proposed method can accurately extract the feature information from the

original vibration signals, and has higher diagnosis accuracy. SSA helps to optimize the parameters and improve the classification performance of SVM. In the future, we will use data from other industries and scenarios for diagnosis, and further investigate the improvement of model performance and diagnostic accuracy.

ACKNOWLEDGMENT

This research was supported by Key Scientific and Technological Project of Henan Province (222102210189, 212102210324, 202102210286), Industry-university Cooperative Education Program of Ministry of Education, (202002029035), Key Scientific Research Projects of the Higher Education Institutions of Henan Province (22B460004), and Doctoral Fund of Henan Institute of Technology (KY1750).

REFERENCES

- [1] A.J. Guillén, A. Crespo, M. Macchi, and J. Gómez, "On the role of prognostics and health management in advanced maintenance systems," *Production Planning & Control*, vol. 27, pp. 991-1004, 2016. [Online]. Available: <https://doi.org/10.1080/09537287.2016.1171920>.
- [2] H.O.A. Ahmed, M.L.D. Wong, and A.K. Nandi, "Intelligent condition monitoring method for bearing faults from highly compressed measurements using sparse over-complete features," *Mechanical Systems and Signal Processing*, Vol. 99, pp.459-477,2018. [Online]. Available: <https://doi.org/10.1016/j.ymssp.2017.06.027>.
- [3] J. P. Gómez, F. E. Hernández Montero, J. C. Gómez Mancilla and Y. V. Rey, "Identification of Babbitt Damage and Excessive Clearance in Journal Bearings through an Intelligent Recognition Approach" *International Journal of Advanced Computer Science and Applications*(IJACSA), vol. 12, pp. 526-533, 2021. [Online]. Available: <http://dx.doi.org/10.14569/IJACSA.2021.0120467>.
- [4] S. Schmidt, P. S. Heyns, and K. C. Gryllias, "A pre-processing methodology to enhance novel information for rotating machine diagnostics," *Mechanical Systems and Signal Processing*, vol. 124, pp. 541-561, 2019. [Online]. Available: <https://doi.org/10.1016/j.ymssp.2019.02.005>.
- [5] Y. Wei, Y. Li, M. Xu, and W. Huang, "A review of early fault diagnosis approaches and their applications in rotating machinery," *Entropy*, vol. 21, article number: 409, 2019. [Online]. Available: <https://doi.org/10.3390/e21040409>.
- [6] L. Ruonan, Y. Boyuan, Z. Enrico, and C. Xuefeng, "Artificial intelligence for fault diagnosis of rotating machinery: A review," *Mechanical Systems and Signal Processing*, vol. 108, pp. 33-47, 2018. [Online]. Available: <https://doi.org/10.1016/j.ymssp.2018.02.016>.
- [7] D. Stuaní Alves et al., "Uncertainty quantification in deep convolutional neural network diagnostics of journal bearings with ovalization fault," *Mechanism and Machine Theory*, vol. 149, article number: 103835, 2020. [Online]. Available: <https://doi.org/10.1016/j.mechmachtheory.2020.103835>
- [8] M. Altaf , T. Akram, M. A. Khan, M. Iqbal, M. Ch, and C. H. Hsu, "A New Statistical Features Based Approach for Bearing Fault Diagnosis Using Vibration Signals," *Sensors*. Vol. 22, article number: 2012, 2022. [Online]. Available: <https://doi.org/10.3390/s22052012>.
- [9] Iqbal, M., Madan, A.K., "CNC Machine-Bearing Fault Detection Based on Convolutional Neural Network Using Vibration and Acoustic Signal," *Journal of Vibration Engineering & Technologies*, 2022. [Online]. Available: <https://doi.org/10.1007/s42417-022-00468-1>.
- [10] X. J. Shi, J.K. Zhang, H. Du, and J.C. Zhang, "An Experimental Study of Sensor-less Fault Diagnosis on Rolling Bearing of Wind Turbine Generator System," *Proceedings Of The 2015 International Power, Electronics And Materials Engineering Conference*, vol. 17, pp. 422-426, 2015.
- [11] Y. Cheng, Z. Wang and W. Zhang, "A Novel Condition-Monitoring Method for Axle-Box Bearings of High-Speed Trains Using Temperature Sensor Signals," *IEEE Sensors Journal*, vol. 19, pp. 205-

- 213, 2019. [Online]. Available: <https://doi.org/10.1109/JSEN.2018.2875072>.
- [12] P. J. Hyun, and C. H. Kim, "Analysis of Accuracy and Computation Complexity of Bearing Fault Diagnosis Methods using CNN-based Deep Learning," *The Journal of Korean Institute of Next Generation Computing*, vol. 18 pp. 7-18, 2022. [Online]. Available: <https://doi.org/10.23019/kingpc.18.1.202202.001>.
- [13] R. S. Gunerkar, A. K. Jalan, and S. U. Belgamwar, "Fault diagnosis of rolling element bearing based on artificial neural network," *Journal of Mechanical Science and Technology*, vol. 33, pp. 505-511, 2019. [Online]. Available: <https://doi.org/10.1007/s12206-019-0103-x>.
- [14] X. M. Yao, S. B. Li, J. J. Hu, "Improving Rolling Bearing Fault Diagnosis by DS Evidence Theory Based Fusion Model", *Journal of Sensors*, vol. 2017, article number: 6737295, 2017. [Online]. Available: <https://doi.org/10.1155/2017/6737295>.
- [15] K. H. Zhu, L. Chen, X. Hu, "Rolling element bearing fault diagnosis based on multi-scale global fuzzy entropy, multiple class feature selection and support vector machine," *Transactions of the Institute of Measurement and Control*, vol. 41, pp. 4013-4022, 2019. [Online]. Available: <https://doi.org/10.1177/0142331219844555>.
- [16] A. Kablaa and K. Mokrani, "Bearing fault diagnosis using Hilbert-Huang transform (HHT) and support vector machine (SVM)," *Mechanics and Industry*, vol. 17, article number: 308, 2016. [Online]. Available: <https://doi.org/10.1051/meca/2015067>.
- [17] M. M. Song, H. X. Song, and S. G. Xiao, "A study on Fault Diagnosis Method of Rolling Bearing Based on Wavelet Packet and Improved BP Neural Network," *IOP Conference Series: Materials Science and Engineering*, vol. 274, article number: 012133, 2017. [Online]. Available: <https://doi.org/10.1088/1757-899X/274/1/012133>.
- [18] J. Liu, and X. Yu, "Rolling Element Bearing Fault Diagnosis for Complex Equipment based on MFMD and BP Neural Network," *Journal of Physics Conference Series*, vol. 1948, article number: 012133, 2021. [Online]. Available: <https://doi.org/10.1088/1742-6596/1948/1/012113>.
- [19] Q. E. Zhao, H. W. Huang, and K. Feng, "Mixed Fault Diagnosis for Rolling Bearings Based on Wavelet Packet Energy- Decision Tree," *Bearing*, vol. 6, pp. 43-46, 2016. [Online]. Available: <https://doi.org/10.3969/j.issn.1000-3762.2016.06.011>.
- [20] T. Xia, Y. Zhan, and J. B. Guo, "Bearing fault diagnosis based on Wavelet Packet and Gradient Boosting Decision Tree," *Journal of Shaanxi University of Science & Technology*, vol. 38, pp. 144-149, 2020. [Online]. Available: <https://doi.org/10.3969/j.issn.1000-5811.2020.05.023>.
- [21] Y. J. Chen, H. Yuan, H. M. Liu, and C. Lu, "Fault diagnosis for rolling bearing based on SIFT-KPCA and SVM," *Engineering Computations*, vol. 34, pp. 53-65, 2017. [Online]. Available: <https://doi.org/10.1108/EC-01-2016-0005>.
- [22] S. N. Chegini, A. Bagheri, and F. Najafi, "A new intelligent fault diagnosis method for bearing in different speeds based on the FDAF-score algorithm, binary particle swarm optimization, and support vector machine," *Soft Computing*, vol. 24, pp. 10005-10023, 2020. [Online]. Available: <https://doi.org/10.1007/s00500-019-04516-z>.
- [23] M. F. Zheng, and H. Y. Quan, "Rolling bearing fault diagnosis of SVM optimized by surface-simplex swarm evolution," *Journal of Chongqing University(Natural Science Edition)*, vol. 44, pp. 43-52, 2021. [Online]. Available: <https://doi.org/10.11835/j.issn.1000-582X.2020.278>.
- [24] T. Wu, C. C. Liu, and C. He, "Fault Diagnosis of Bearings Based on KJADE and VNWOA-LSSVM Algorithm," *Mathematical Problems in Engineering*, vol. 2019, article number: 8784154, 2019. [Online]. Available: <https://doi.org/10.1155/2019/8784154>.
- [25] C. L. Zhang, and S. F. Ding, "A stochastic configuration network based on chaotic sparrow search algorithm," *Knowledge-Based Systems*, vol. 220, article number: 106924, 2021. [Online]. Available: <https://doi.org/10.1016/j.knsys.2021.106924>.
- [26] X. Yan, and M. Jia, "A novel optimized SVM classification algorithm with multi-domain feature and its application to fault diagnosis of rolling bearing," *Neurocomputing*, vol. 313, pp. 47-64, 2018. [Online]. Available: <https://doi.org/10.1016/j.neucom.2018.05.002>.
- [27] P. Li, "Fault Diagnosis of Motor Rolling Bearing Based on GWO-SVM," *International Core Journal of Engineering*, vol. 5, pp. 238-245, 2019.
- [28] Z. Guo, M. Liu, Y. Wang, and H. Qin, "A New Fault Diagnosis Classifier for Rolling Bearing United Multi-scale Permutation Entropy optimize VMD and Cuckoo Search SVM," *IEEE Access*, vol. 8, pp. 153610-153629, 2020. [Online]. Available: <https://doi.org/10.1109/ACCESS.2020.3018320>.
- [29] Y. Wang, C. N. Xu, Y. Wang, and X. Z. Cheng, "A comprehensive diagnosis method of rolling bearing fault based on CEEMDAN-DFA-improved wavelet threshold function and QPSO-MPE-SVM," *Entropy*, vol. 23, article number: 1142, 2021. [Online]. Available: <https://doi.org/10.3390/e23091142>.
- [30] J. K. Xue, and B. Shen, "A novel swarm intelligence optimization approach: sparrow search algorithm," *Systems Science and Control Engineering*, vol. 8, pp. 22-34, 2020. [Online]. Available: <https://doi.org/10.1080/21642583.2019.1708830>.
- [31] W. A. Smith, R. B. Randall, "Rolling element bearing diagnostics using the Case Western Reserve University data: A benchmark study," *Mechanical Systems and Signal Processing*, vol. 64-65, pp. 100-131, 2015. [Online]. Available: <https://doi.org/10.1016/j.ymssp.2015.04.021>.

Multi-instance Finger Vein-based Authentication with Secured Templates

Swati K. Choudhary¹

Department of Electronics Engineering
K. J. Somaiya College of Engineering
Mumbai, India

Ameya K. Naik²

Department of Electronics and Telecommunication
Engineering, K. J. Somaiya College of Engineering
Mumbai, India

Abstract—The illegitimate access to biometric templates is one of the major issues to be handled for authentication systems. In this work, we propose to use two instances of finger vein images which inherits the advantages of a robust multi-modal biometric authentication system without needing different sensors. Two local texture feature extraction methods are experimented on standard finger-vein datasets. Fused discriminating features with reduced dimension lowers down the system computational cost. A cancelable template protection scheme as Gaussian Random Projection based Index-of-Max is then applied for embedding privacy and security to the templates. Foremost template protection properties like revocability, non-invertibility and unlinkability are observed to be significantly obeyed by the proposed system with considerable authentication performance. Recognition performance of the proposed methods are compared with some previously executed finger vein systems and observed to be less complex and overperforming on the combined basis of authentication and template protection. Thus, the proposed system utilizes multiple evidence and provides a balanced performance with respect to authentication, template protection and computational cost.

Keywords—Finger vein; multi-instance; authentication; cancelable; template protection

I. INTRODUCTION

Information Technology and Internet driven life has created the extreme need for securing the evidence related to personal identity. Biometrics based authentication systems put up some problems related to security and privacy of data [1], owing to which an efficient template protection scheme needs to be employed. The requirements regarding irreversibility, revocability, unlinkability and performance preservation should be satisfied by an effective template protection technique. To meet these requirements, various techniques have been investigated as bio-cryptosystems and cancelable biometrics. Among them, Cancelable biometrics is more appropriate technique to handle both, the security and privacy of templates by repeatedly deforming the template features using some transformation.

The design of transformation function should be extremely difficult to invert, enabling computational infeasibility to get the original features back from the transformed template. This characteristic prevents the privacy attack to get original data back. Renewability is another property that the designed transformation function should come up with. This enables generation of newly transformed template if the previously

enrolled template is compromised. Additionally, templates created by using different transformations should not match with each other applying diversity to the template protection technique. Furthermore, it should be challenging to differentiate between the templates established by same biometric information. This prevents cross matching of biometric data across various applications, obeying the unlinkability property. More importantly, when all the above-mentioned security and privacy preserving characteristics are considered to design a transformation function, it should not reduce the authentication level. The proposed framework applied the Gaussian Random Projection based Index of Max (GRP-IoM) hashing technique [2] on the real valued finger vein features to generate cancelable and highly non-invertible protected templates.

The proposed work is utilizing multiple instances of finger veins for authenticating people. Finger veins are innate, non-intrusive, and highly resistant to duplication and captured by small imaging sensor. Additionally, distinct patterns for twins [3] and liveness detection property are included in finger veins. Occasionally, finger vein images suffer from illuminations and blood fluctuations and becomes low contrast and unstable. Thus, to overcome this low-quality issue to some extent, combination of multiple instances of finger vein images will be beneficial minimizing the limitations of unimodal biometrics viz. non-universality, intra-class deviations, inter-class resemblance, scope for spoofing, etc. Thus, in this work, we propose fused-discriminant (FD) feature set extraction from two instances of finger vein images with two different local texture-based features. The feature extraction techniques experimented is Uniform Local Binary Pattern (ULBP) [4] and Local Hybrid Binary Gradient Contour (LHBGC) [5]. The texture features from the two finger vein instances are fused as well as reduced in size by maintaining their correlation within an individual and enhancing their discrimination between the individuals. This improves authentication performance and reduces computational intricacy transformation for template protection and matching. Thus, the key contributions of the proposed framework are outlined as follows:

- Generating highly irreversible, un-linkable and revocable templates by using fused-discriminant feature set.
- Person authentication is based on two separate identity proofs rather than single finger vein image to overcome the issues like low quality with lower computation cost.

- Two different texture-based features (ULBP and LHBGC) are experimented in their fused-discriminant forms for generating protected templates and preserving authentication performance, significantly.

This paper is organized as follows. Section II describes some of the previous work carried on, in the field of single and multi-instance finger vein authentication and a variety of template protection methods that has been practiced for the same. Section III provides description regarding the fundamental methods used in the proposed authentication system with template protection. Thereafter, in Section IV, details of implementation process for the proposed multi-instance finger vein-based authentication system and obtained experimental results for authentication and template protection are represented and analyzed. Validations of obtained authentication and time complexity results are shown by comparing with some earlier reviewed work. To conclude, Section V covers the overall contribution of the proposed authentication framework and its further scope.

II. RELATED WORK

Various approaches have been practiced for utilizing finger veins for authentication purpose. Handling the security issue of biometric templates with authentication is a great challenge. In Section II-A, different methodologies implemented for finger vein-based authentication are discussed. Further, Section II-B gives overview of various schemes practiced for protecting finger vein templates.

A. Finger Vein based Systems

There are line-based, point-based and texture-based features popularly used for vascular pattern finger vein regions. Line-based finger vein feature extraction is initiated by [3] in the form of Repeated Line Tracking (RLT) method through a line tracking algorithm. The author continued his research in [6] by extracting the center lines of the veins, calculating their local maximum curvatures. This method is found to be robust against variations in vein-widths. Later, [7] proposed Wide Line Detector (WLD) with comparatively faster vein pattern extraction but at the expense of degradation in authentication performance. In [8], Enhanced Maximum Curvature (EMC) method is used for feature extraction which identifies fine delineations in vein-patterns using Histogram of Oriented Gradients (HOG) but observed to be slower than WLD. Alternatively, some point-based feature extraction methods are also practiced as minutiae points in [9] or multiple key point sets from SIFT (Scale-invariant Feature Transform) in [10]. Ultimately, point based features also needs vein patterns involving computational cost for extracting vein pattern. In [11], special points as cross/end points of veins and connections between them are matched to reduce the finger vein matching time. The performance of this scheme is sensitive to Region of Interest (ROI) and used only good quality images for evaluation.

In texture-based approach, various classical and innovative methods for extracting texture information have been practiced. Most widespread method adept of extracting recognizable features from such images is Local Binary Pattern (LBP). This LBP method introduced by [12] is fast in which

texture features are obtained from gray level difference in neighborhood pixels. These classical LBP features are lengthy and observed to be very sensitive to image translations and rotations. To cope with this, there are various LBP variants proposed previously for finger veins and other biometrics [13]. In [14], t-norm based fusion of LBP feature scores belonging to two finger vein instances is implemented using hamming distance. Accurate authentication primarily depends on discriminability of features. Another variant of LBP as Local Hybrid Binary Gradient Contour (LHBGC) features [5] are shown to be more informative on finger veins compared to [13, 14]. This method computes local histograms to compute the frequencies of sign and magnitude components, locally in the image. Further, Uniform LBP (ULBP) texture feature extraction proposed by [4] is very much suitable to finger vein structure. As the vein will either lie inside or cross the neighborhood in vein-region, the resulting pattern will not have many distinguished bitwise transitions. Thus, these ULBP features which are more compact than LBP, covering majorly uniform patterns are suitable for finger vein. ULBP features preserving spatial information are found to have some degree of invariance to rotation, pose and illumination because of histogram computation over image partitions. Variety of feature extraction methods complementing with different classifiers have been practiced for finger vein images [15-20]. Some recent finger vein-based authentication implemented Machine Learning (ML) and Artificial Intelligence (AI) approach especially for identification, if huge data needs to be worked on. [21] utilized VGG-Net-16, which is composed of thirteen convolutional fully connected layer model finely tuned and pre-trained with two finger vein image difference. A deep learning-based technique is proposed by [22] to work on finger vein images of varying quality. This network comprised of four convolutional layers and is experimented on four different databases. Deep learning methods requires heavy processing configuration with huge amount of training data. Parameter tuning is another complex process to be handled in case of AI applications.

In this work, we offer to use two variants of LBP as Uniform LBP (ULBP) and Local Hybrid Binary Gradient Contour (LHBGC), considering their feature discriminability and suitability for finger veins. Additionally, LBP based methods are proven to be resistant to uneven shading and saturation from input imaging devices [23]. The proposed authentication is using ULBP and LHBGC features for multiple instances of finger vein, in their Fused-Discriminant form as FD-ULBP and FD-LHBGC respectively for reducing their dimensions and enhancing discriminability. These features are further transformed for template protection and their authentication performances are analyzed in transformed domain.

B. Finger Vein Template Protection

Severe security and privacy threats are faced by biometric templates stored in database, regarding unauthorized access, original feature breeding from coded template or from cross matching across applications. Bio-cryptosystems (BCS) and cancelable biometrics (CB) are the major template protection schemes employed to secure the templates. As the proposed framework is implementing CB for the previously mentioned

advantages, major emphasis is given on CB techniques. CB schemes are enormously practiced to secure various popular unibiometric traits as face [24], fingerprint [25] and iris [26]. CB techniques for different combinations of multi-biometric traits have also been practiced. Fusion of multiple traits at different levels are investigated in [27] for better performance. Feature level fusion [28] is advantageous over others, especially if template protection is needed as single fused protected template has to be handled. On the contrary, compatibility issues for features generated out of different biometric traits is highly challenging. When these fused-diverse features are transformed to form cancelable templates, preserving recognition performance is critical. A balanced solution to this is multi-instance biometrics offering multiple evidence-based authentications with compatible features and no extra sensor.

Focusing on CB for securing finger vein images [29] proposes a combined CB and BCS approach by applying cancelable bio-hashing method to finger vein image. Bio-hashing transforms finger vein Gabor features compacted by Linear Discriminant Analysis (LDA) into binary string. Then the Fuzzy Commitment Scheme (FCS) and Fuzzy vault is applied to binary string. In [30], a similar approach of random projection based cancelable transform is applied on Gabor features reduced by Principal Component Analysis (PCA) and then FCS is operated on the resulted binary feature set. L2-Norm is used to classify the transformed templates. [31] again offers a random projection based cancelable transformation of vein end and intersection points and classification is done using Deep Belief Network (DBN). This approach requires password and a huge dataset for training. In [32], Bloom filter template security is utilized for fingerprint, signature, and their fused features. As Bloom filter technique can be applied to fixed length binary features, some fingerprint-oriented methods like minutiae cylindrical code employed to finger vein [33] can be utilized for Bloom-filter based CB. [34] proposed CB for finger vein image through block-remapping and image-wrapping based transformation before feature extraction in image domain. Gabor features are used for verification and renewability evaluation. Block remapping in image domain is observed to provide better performance than wrapping which is also dependent on block size. Recently, [35] proposed CB method using on Index of Max [2] with alignment robust hashing (ARH). This work has also experimented with block-remapping and image wrapping but in feature domain to extract binary features. In [35], various binary feature extraction methods are experimented with three different types of CB techniques. ARH with Index of Max hashing works fine for alignment-free situations but inferior to image wrapping in feature domain with respect to authentication performance. Also, there is no clear empirical analysis of revocability for ARH-Index of Max hashing. AI-based methods like Convolution Neural Network (CNN) are very popular for identification but needs huge data volume. Also, as CNN is reversible in nature, the original raw features fed to CNN for classification can be inverted back. In [36], finger vein bio-hashed binary features are transformed into non-invertible and renewable code using Binary Decision Diagram (BDD). This protected code is fed to the Multi-Layer Extreme Machine

Learning for verification and identification which runs the system fast. Recently, [37] also implemented bio hashing for securing finger vein templates made out of deep features using multi-term loss function showing excellent verification outcome but have not investigated for template protection factors like unlinkability or irreversibility.

In this article, we propose to use Gaussian Random Projection-based Index of Max (GRP-IoM) hashing for non-linear mapping of real valued fused feature set from instances of finger vein to generate integer based protected templates. Authentication performances along with corresponding computational costs of the proposed framework with two different texture features are evaluated. Theoretical and empirical evaluation of non-invertibility, unlinkability and revocability for the protected finger vein templates is also provided.

III. PROPOSED METHODOLOGY

The proposed framework is for person authentication using cancelable transformation of multiple instances of finger veins. Particularly, vascular patterns composed of veins within the human physique are difficult to counterfeit, contactless and concurrently provides liveness detection. Multi-instance finger vein-based authentication system makes the overall recognition performance depend on a greater number of biometric facts. Cancelable and distorted versions of fused features from multiple finger veins creates a secure verification system based on intrinsic biometric trait. This ensures renewability/revocability of our permanent unique features which are quite un-linkable across various applications. Various methods involved in this combined approach of authentication and template protection are explained as follows:

A. Finger Vein Processing

Finger vein verification involves challenges as finger vein images suffer from illumination and blood fluctuation factors. Captured finger vein images with poor contrast and misalignment may deteriorate authentication performance. So, for reliable finger vein verification, proper pre-processing and feature extraction techniques are needed. Details of methods used in the proposed framework for finger vein pre-processing and feature extraction are explained below:

1) *Finger vein pre-processing*: Finger vein pre-processing involves Region of Interest (ROI) extraction with finger alignment and image enhancement. ROI is localized by firstly detecting finger outline by edge detection algorithm and fitting a center line into the finger. The finger is gradually rotated and shifted using this center line until it gets aligned in the middle of the image with horizontal posture. The rectangular ROI is extracted masking out the background portion. Further, adaptive histogram equalization technique is used on finger vein ROI images to improve the vein pattern visualization. The noise removal is done using Weiner filter. Fig. 1 shows the overall result images for finger vein pre-processing in a sequential manner. For further details on preprocessing please refer to [38].

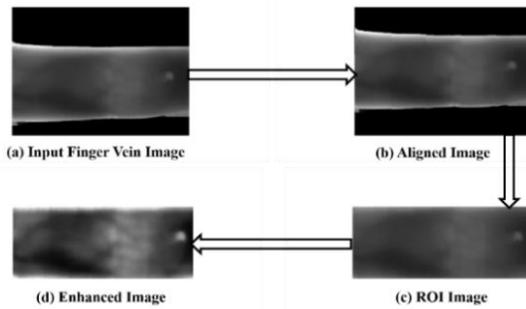


Fig. 1. Finger Vein Pre-processing.

2) *Finger vein feature extraction*: In the proposed approach, texture features belonging to two finger vein instances are combined with enhanced discriminability and compaction and then used for person authentication. Further these fused feature set is transformed for generating cancelable protected templates to incorporate template security. Two types of texture-based features are experimented, namely histograms of Uniform Local Binary Pattern (ULBP) and Local Hybrid Binary Gradient Contour (LHBGC). Both features are variants of original LBP method which is computationally less complex and extracts fine scale textures [12].

The first method implemented for finger vein texture feature extraction is histogram of uniform LBP. Original LBP description of a pixel in its canonical form is simple and created by comparing the intensity of P neighboring pixels to the center pixel in some radius r . These P comparisons in clockwise or anticlockwise direction is interpreted as a binary vector. Binary 1 is taken if the center pixel intensity is less than neighboring pixel, otherwise binary 0 is taken. This $LBP_{P,r}$ feature extraction process replaces each pixel in original image to a binary pattern except border pixels which do not have all neighboring pixels. These feature vectors can be transformed into histogram with $2P$ bins as each conceivable LBP is assigned to an individual bin. Various combinations of r and P are practiced but the most popular combination is $P=8$ with $r=1$. Proposed framework is also using $P=8$ with $r=1$ focusing on histogram dimensionality for computation, memory consumption and minimal original information loss. Further, it has been observed that specific binary patterns count for fundamental texture properties known as “uniform” patterns [4]. These uniform LBP features are observed to carry gray scale invariance and rotation invariance, but additional computations for rotation invariance are not incorporated in the present work as alignment of finger vein pattern is implemented in pre-processing. LBP is termed as uniform if it consists of maximum two 1-0 or 0-1 switching, viewing the bit pattern in circular way. For example, the pattern 00000110 is uniform whilst 10100000 is not. It is noticed that uniform patterns accounted for approximately 90 percent of all patterns. Hence, least information was lost by handing over all non-uniform patterns to one non-uniform category. If practiced particularly for $P=8$, it is seen that just 58 of the 256 possible 8-bit strings are uniform, we can thus encode all 256, 8-bit local binary patterns using 59 (58 uniform and 1 non-uniform)

codewords. To indicate that uniform patterns are being used, $u2$ is added as superscript to the LBP operator to generate $LBP_{P,r}^{u2}$. Selecting the uniform patterns thus reduces the histogram length to 59 ($P*(P-1) + 3$) bins from 256 ($2P$) bins for $P=8$. Histograms of individual patterns show lower distinguishability as compared to that of uniform patterns accounting for the variations in their statistical properties.

Fig. 2 explains the overall process of uniform LBP feature extraction. Each input pre-processed image instance of finger vein is firstly split into N number of blocks. In the present work, each block is of size 8×8 with $P=8$ and $r=1$. Histograms were then computed for each block and the resulting histograms were then concatenated together to form one feature vector. Spatial information was implicitly encoded into this feature vector from the order in which the histograms were concatenated. Hence, after histograms of uniform LBP feature calculation the overall combined histogram feature length for the entire image is of length, $59 \times N$. These texture features for the two finger vein image instances are calculated and further processed for feature fusion with dimension reduction and template protection.

The second method experimented for feature extraction of finger vein images is Local Hybrid Binary Gradient Contour (LHBGC) [5], which is also an LBP variant. LHBGC features are considered for its property of extensive information content regarding finger vein authentication as compared to various other texture features. This method computes the local histograms counting for frequencies of sign and magnitude for finger vein images, locally. Firstly, the preprocessed image of finger vein is subjected to sign and magnitude component extraction followed by local histogram calculation. Fig. 3 shows sign and magnitude computation in 3×3 neighborhood periphery for an input image. Adjacent pixel intensities are compared in 3×3 neighborhood periphery and these distances, b_i for $i=0, 1, \dots, 7$ are calculated using Eq. (1) which are further break down into sign and magnitude components. Hence, sign ($[s_0, s_1, \dots, s_7]$) and magnitude ($[m_0, m_1, \dots, m_7]$) vectors are obtained by decomposing distances ($[b_0, b_1, \dots, b_7]$) as shown in Eq. (2). The sign values are equivalent to basic Binary Gradient Contour codes and this binary code is further translated to decimal number. The magnitude component value for each 3×3 region is calculated by adding each of $[m_0, m_1, \dots, m_7]$.

$$b_i = p_i - p_{(i+1) \bmod 8}, i = 0, 1, 2, \dots, 7 \quad (1)$$

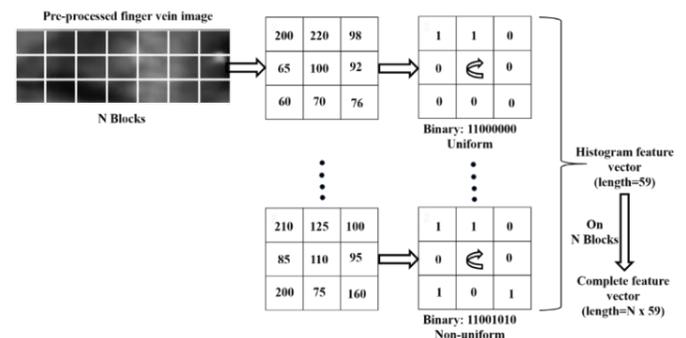


Fig. 2. Uniform Local Binary Pattern Histogram Feature Extraction Process.

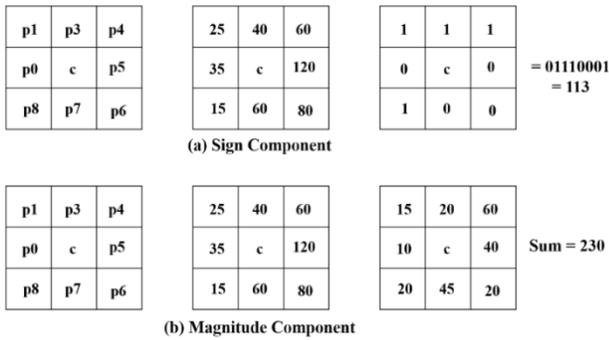


Fig. 3. Sign and Magnitude Component Calculation for Local Hybrid Binary Gradient Contour (LHBGC) Features.

where, $[p0, p1, \dots, p7]$ are the adjoining pixels along the periphery of 3×3 neighborhoods.

$$b_i = s_i * m_i \text{ and } \begin{cases} s_i \\ m_i = |b_i|, s_i = \begin{cases} 1, b_i \geq 0 \\ -1, b_i < 0 \end{cases} \end{cases} \quad (2)$$

The sign and magnitude components are distributed equally into number of cells for local histogram computation. For every single cell, a local histogram is computed, and the sign histogram bin is voted in a biased way by each pixel of the magnitude component in that cell based on the value present in the sign component. For further details on local histogram computation, please refer [5]. Various parameters involved in LHBGC feature extraction process are number of cells i.e., number of rows and columns with number of bins. The local histograms belonging to every cell are formed as vectors which are all concatenated to produce a combined feature vector for a finger vein image. This concatenated feature vector is of high dimension. Thus, further processing on this basic feature set is proposed for reducing feature dimensionality with enhanced feature discriminability.

3) *Finger vein feature fusion*: In the proposed approach, two texture-based features of finger vein image are experimented, namely ULBP and LHBGC as explained in the above section. The present work contributes for the effective person authentication using two instances of finger vein by fusing them into compact and informative feature set. Both ULBP and LHBGC histogram features are extracted from right index (RI) and right middle (RM) finger vein instances and fused using Discriminant Correlation Analysis (DCA) to form their fused-discriminant versions as FD-ULBP and FD-LHBGC. DCA reduces fused feature set dimensionality considerably (number of subjects present in training set), implemented by summing up the individual finger vein discriminant feature vectors. DCA is an effective tool for fusing features in pattern recognition. It is computationally efficient and applicable in real-time situations [39]. DCA obtains data from multiple feature sets and incorporates the class structure into canonical correlation analysis via transformations, thus taking into account the differences in the different classes while at the same time maximizing the pairwise correlations among the features in the two feature sets.

More details about how DCA works can be found in [39]. The proposed work contributes by authenticating person based on deformed version of compact and fused feature set for the two-finger vein instance evidence. The deformation of feature set is needed for template privacy and security.

B. Finger Vein Template Protection and Matching

Finger vein features are protected by creating their hashed codes using randomly generated projection matrices. These codes are observed as highly irreversible and revocable in nature. The overall process is known as Gaussian Random Projection-based Index of Max (GRP-IoM) and is proposed by [2] for uni-biometric fingerprint system. This hashing technique is implemented for multiple instances of finger vein images in the proposed system.

The fused feature set of multi-instance finger veins as $x \in R^d$ is projected onto a d -dimensional random Gaussian vectors as $k \in R^d$. The overall hashing process is operated as follows:

- 1) Generate q number of d -dimensional Gaussian random vectors as k_1, k_2, \dots, k_q and form a projection matrix, $W^i = [w^1, w^2, \dots, w^q]$.
- 2) Project input fused feature vector, x onto W^i and record index of maximum value from $\Phi_i(x) = \arg \max_{i=1,2,\dots,q} (W^i, x)$ as t .
- 3) Repeat steps 1 and 2, n number of times to obtain hashed feature set as $t = (t_1, t_2, \dots, t_n)$.

The IoM hashing basically obey the ranking based locality sensitive hashing that attempts to confirm that any two highly similar feature vectors result to greater probability of collision. On the other side, the dissimilar vectors result in smaller probability of collision. Assuming the collision probability of two hashed codes as enrolled template, $t^e = \{t_j^e | j=1, \dots, n\}$ and query template, $t^q = \{t_j^q | j=1, \dots, n\}$ is represented as $CP[t_j^e, t_j^q] = ss(t^e, t^q)$ for $j=1, 2, \dots, n$. Thus, the higher value of $ss(t^e, t^q)$ signifies high probability of collision. Computationally, $ss(t^e, t^q)$ is observed as the number of zeroes (collisions) counted in subtracting t^e and t^q , element-wise for n number of iterations which is considered as a template match score. In case of compromised template, IoM hashed code will be renewed by generating new random Gaussian matrices.

C. Multi-instance Finger Vein based Secured Authentication

The entire application of the proposed framework for multi-instance finger vein based biometric verification system with template protection is categorized into two major parts as enrollment and authentication. As shown in Fig. 4, enrolling an individual in the system involves finger vein pre-processing, feature extraction, feature fusion from multiple instances and creation of cancelable protected templates for secured authentication. Fig. 1 depicts pre-processing which involves alignment of individual finger vein images, ROI selection and image enhancement. Fig. 4 describes the overall enrolment and authentication process of right index and right middle finger vein images with feature extraction, feature fusion and cancelable protected templates.

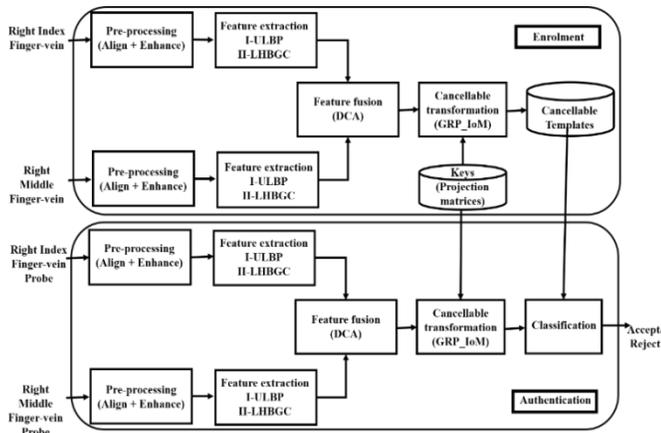


Fig. 4. Block Diagram of Proposed Multi-instance Finger Vein-based Authentication.

In the proposed work, two localized texture-based feature extraction methods are experimented on finger vein images to extract features namely, method-I, Local Hybrid Binary Gradient Contour (LHBGC) and method II, Uniform Local Binary Pattern (ULBP). Both, LHBGC and ULBP feature extraction techniques are explained in section III-A.2. Features extracted from multiple instances, particularly from right index (RI) and right middle (RM) finger vein images are then fused to create a combined feature set using Discriminant correlation analysis (DCA) [39]. DCA feature fusion is practiced with summation technique which also involves feature dimension reduction to significant extent. Thus, the fused-discriminant feature set for RI and RM finger vein instances as FD-LHBGC and FD-ULBP are ready for cancelable transformation meant for template protection.

Fused finger vein feature set is then converted to hashed code using Gaussian Random Projection based Index of Maximum (GRP-IoM) hashing technique [2]. This generic hashing scheme is highly irreversible and un-linkable cancelable transform. The irreversibility and un-linkability of the proposed template protection method on fused multi-instance finger vein feature set is shown in experimentation section. Finally, the cancelable protected templates are stored in database with their corresponding projection matrices (keys).

Authentication of a person also requires pre-processing, feature extraction and hashed code transforms to be generated in the same manner as that during enrolment. Classification of an identity is carried on using collision classifier as explained in Section III-B. Authentication and template protection evaluations for the proposed framework are demonstrated and analyzed in the experimentation section.

IV. EXPERIMENTAL RESULTS AND DISCUSSION

Experiments have been conducted on the real multi-instance biometric datasets to evaluate the performance of the proposed multi-instance finger vein-based authentication methodology with respect to verification performance and template protection.

A. Database

Two openly available standard databases are used for evaluating the overall results of the proposed multi-instance finger vein-based verification system as SDUMLA-HMT and UTFVP. The SDUMLA-HMT (Shandong University Machine Learning and Applications) biometric database [40] consists of finger vein images for 106 individuals in addition to samples of additional biometric traits such as the face and iris. There are vein images of six fingers (three on each hand) and since vascular patterns differ on every finger of every hand, there are effectively $106 \times 6 = 636$ possible classes to identify. Furthermore, six images per class were captured resulting in a total of 3816 images. Each image is 320 x 240 pixels in dimension and stored in the uncompressed, bitmap file format. The UTFVP finger vascular pattern database [41] was produced by University of Twente, Nederland. It consists of 1440 images from 60 persons with 4 images per instance of 6 different fingers. The resolution of each image is 672 x 380 pixels stored in 8-bit gray scale PNG format.

B. Experimentation

We ensured that the data used to train the feature fusion process was never used to test it so that a true indication of its predictive accuracy could be obtained. For SDUMLA-HMT database, a total of 1272 images were used in carrying out the experiments. Individually for 106 subjects, four samples are used for training with DCA which executes feature fusion and dimension reduction and two samples for authentication testing. Combination of right middle (RM) and right index (RI) finger veins is considered for experimentation as it is the most popular and convenient choice for multi-instance system. Experimentation is carried on with total of 212 (106 x 2) genuine scores and 44520 ((212-2) x 212) imposter scores for SDUMLA-HMT database. Similarly, in case of UTFVP database, total 480 finger vein images are used for experimentation. Training involved two samples per class and testing is done on other two samples for multiple instances. Experimentation is done on UTFVP with total of 120 (60x2) genuine and 14160 (120 x (120-2)) imposter scores. All the experiments are implemented using MATLAB 2019a on a system i5-CPU with 2.5 GHz and 4 GB memory.

1) *Verification performance evaluation:* The verification performance of the proposed multi-instance finger vein authentication system is evaluated in terms of percentage equal error rate (EER). Fig. 5 shows Receiver Operating Characteristics (ROC) curves for the proposed verification systems, which also indicates the corresponding EER i.e., the error at where false acceptance rate (FAR) is equal to genuine acceptance rate (GAR). FAR, FRR (false rejection rate) and GAR are represented by following Eqs. (3-5).

$$FAR = \frac{\text{Number of falsely accepted imposters}}{\text{Total number of imposter trials}} \quad (3)$$

$$FRR = \frac{\text{Number of falsely rejected genuines}}{\text{Total number of genuine trials}} \quad (4)$$

$$GAR = 1 - FRR \quad (5)$$

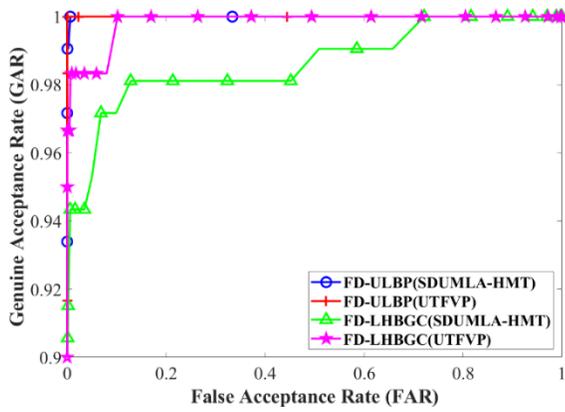


Fig. 5. Receiver Operating Characteristic (ROC) Curves for Proposed Verification with (a) LHBGC and (b) ULBP Feature Extraction.

These ROC curves in Fig. 5 demonstrate that verification performance of the proposed multi-instance (RM+RI) finger vein system is much better with FD-ULBP features as compared to that of FD-LHBGC features. The proposed verification results in the form of EER with practiced feature extraction methods are depicted in Table I with the corresponding feature lengths for two different databases. Since the feature deformation method used for template protection is based on random projections, the EER is computed by considering the average of repeated twenty different randomly generated key-projections.

TABLE I. VERIFICATION PERFORMANCE OF THE PROPOSED SYSTEM

Method	Feature Extraction Method	Verification Performance (%EER) (95% confidence level)	
		Database	
		SDUMLA-HMT	UTFVP
I	FD-LHBGC ^a	4.2 ± 0.19	1.54 ± 0.11
II	FD-ULBP ^b	0.53 ± 0.161	0.00074 ± 0.00119

^aFused Discriminant-Local Histogram Binary Gradient Contour and ^bUniform Local Binary Pattern.

Thus, it is observed that FD-ULBP features are providing better and exceptional verification as compared to that of FD-LHBGC features regardless of change in database. Also, the proposed system is providing considerable verification with incredibly low feature length. This reduces the working plane complexity as well as memory requirement for template storage.

2) *Template protection evaluation:* In this section, privacy and security of templates are evaluated and analyzed for the proposed multi-instance finger vein authentication system. Privacy analysis implies the practical possibility for a template protection technique to tolerate any attack for regaining the intrinsic feature information. Whereas, to achieve considerable attack complexity against the unlawful access to the template protection system through counterfeit features is termed as template security. Particularly for Renewable biometrics, privacy analysis covers non-invertibility of

templates and Attacks via Record Multiplicity (ARM) whereas brute force attack analysis and ARM are included in security analysis. These individual investigations are described as follows:

a) *Privacy analysis:* Privacy analysis covers the assessment of non-invertibility/irreversibility and ARM for templates. Non-invertibility is the measure of computational toughness in retrieving the original feature set from the hashed coded stream with and/or without GRP-IoM method based random key matrices. Even if the number of random Gaussian vectors (q) for all the number of iterations (n) are known to the adversary, there is no clue available to recover the original real valued feature vector (x) from illegitimately obtained hashed coded templates. This is possible because there is no direct link between the projection matrices (token) and the original feature vector due to Index of Max (IoM) property of hashed code. In this case, to break the template privacy the adversary needs to predict the fused real valued features. Considering the worst scenario, let us assume that the maximum and minimum values of original features are known to the adversary for analyzing the guessing complexity. Considering an actual feature example with the minimum and maximum values for a fused feature set obtained for FD-ULBP features tested on SDUMLA-HMT finger vein instances as -0.3361 and 0.3622 , respectively. Suppose the adversary tries to guess from $-0.3361, -0.3360, -0.3359$ and so on, until the maximum 0.3622 . In this case, the total of 6983 options for predicting in the range of four decimal digit precision as in our execution, four decimal digits exactness is fixed. As guessing possibility of a single feature element is coming as $6983 (\approx 2^{13})$ attempts. Hence, the entire feature vector comprising of 105 elements needs around $2^{13 \times 105} = 2^{1365}$ trials. The guessing possibilities for the single and entire feature vector element for the proposed two feature types are shown in Table II. This guessing process is observed to be computationally infeasible. Moreover, the so called guessed feature set is not the original or raw finger vein features but the fused version of two finger vein instances which further adds on to the feature confidentiality.

TABLE II. GUESSING POSSIBILITY FOR SINGLE FUSED FEATURE AND COMPLETE FUSED FEATURE SET

Database (Feature Extraction)	Minimum value	Maximum value	Possibilities for single feature component	Total possibilities for complete feature set
SDUMLA-HMT (FD-ULBP)	-0.3361	0.3622	$6983 \approx 2^{13}$	$2^{13 \times 105} = 2^{1365}$
UTFVP (FD-ULBP)	-0.5243	0.5603	$10846 > 2^{13}$	$2^{13 \times 59} = 2^{767}$
SDUMLA-HMT (FD-LHBGC)	-0.0178	0.0239	$417 \approx 2^9$	$2^{9 \times 105} = 2^{945}$
UTFVP (FD-LHBGC)	-0.105	0.0198	$303 > 2^8$	$2^{8 \times 59} = 2^{472}$

ARM (Attacks via record multiplicity) is a kind of intrusion to template privacy which tries to reconstruct the original biometric data using numerous forfeited templates with or without parameters and information that linked to the algorithm. Specifically, ARM in IoM hashing is computationally tough for deducing the mathematical value as the saved templates are altered into rank space which are not correlated to the finger vein feature space. Thus, the ARM attack intricacy is the same as the non-invertibility attack possibility presented, formerly.

b) *Security analysis:* Security of biometric templates is in danger when threats like brute force attack or masquerade attack or pre-image attack. For GRP based template protection scheme, with $m=150$, $q=70$, guess intricacy for each entry is greater than 26 (70), as indices of hash code takes values between 1 and 70. Therefore, guess complexity for best performance obtained as for 150 (SDUMLA-HMT) and 300 (UTFVP) entries are greater than 2^{900} and 2^{1800} respectively. This is again computationally infeasible.

c) *Unlinkability analysis:* The unlinkability of the implemented multi-instance template protection scheme is validated by involving the pseudo-genuine scores. The pseudo-genuine scores are generated by matching different fused multi-instance finger vein hash codes of the same individual created by utilizing distinct key projection matrices. The pseudo-imposter scores are computed by matching hashed code templates of different individual created by using different key projection matrices, as explained in section IV-B.3.c With this perspective, the overlapping extent of pseudo-genuine and pseudo-imposter distributions indicates the indistinctive ability of the template generation from same or different users. The un-linkability level contributed by the implemented multi-instance finger vein-based authentication framework is indicated by the difficulty in discriminating these hash coded templates.

Fig. 6 and Fig. 7 demonstrate the distribution of pseudo-genuine and pseudo-imposter scores for both the proposed features as FD-ULBP and FD-LHBGC, experimented on SDUMLA-HMT and UTFVP databases. The pseudo-genuine and pseudo-imposter score plots are observed to be mainly overlapped for both the features. These results show that the IoM codes highly meet un-linkability property for the proposed multi-instance finger vein authentication.

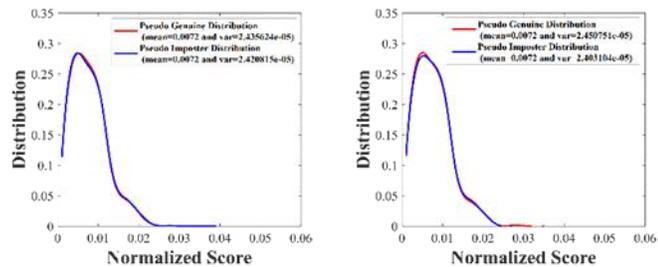


Fig. 6. Un-linkability Analysis using Fused Discriminant Uniform Local Binary Pattern (FD-ULBP) Features on (a) SDUMLA-HMT and (b) UTFVP database.

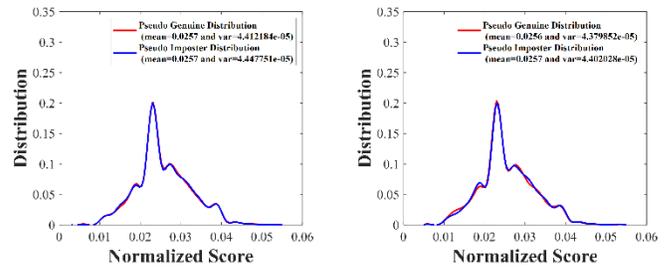


Fig. 7. Unlinkability Analysis using Fused Discriminant Local Hybrid Binary Gradient Contour (FD-LHBGC) Features on (a) SDUMLA-HMT and (b) UTFVP Database.

d) *Revocability analysis:* Revocability or cancelability or renewability is determined by performing the experiments explained in [2]. This generates $105 \times (2 \times 106) = 22260$ and $59 \times (2 \times 60) = 7080$ pseudo-imposter scores for SDUMLA and UTFVP datasets, respectively. The distributions for genuine, imposter and pseudo-imposter scores are exhibited in Fig. 8 and Fig. 9 for both the FD-ULBP and FD-LHBGC features, respectively. It is noticed from Fig. 8 that the imposter and pseudo-imposter distributions for FD-ULBP features are largely overlapped. Fig. 9 implies that the pseudo imposter and imposter scores are not overlapping with each other for FD-LHBGC features, but the pseudo imposter distribution is still distinctive with genuine score distribution preserving differentiation between the two. This entails that even though the hashed codes belonging to the same source finger vein set are freshly created or renewed through newly generated random projection matrices, they are very much distinctive.

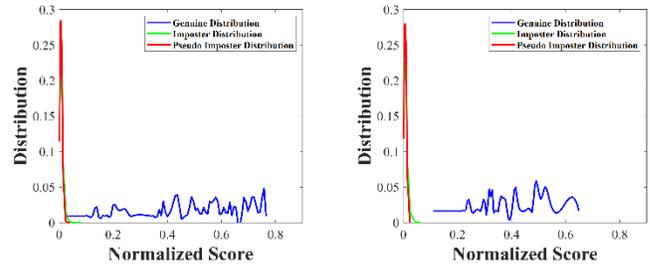


Fig. 8. Revocability Analysis using Fused Discriminant Uniform Local Binary Pattern (FD-ULBP) Features on (a) SDUMLA-HMT and (b) UTFVP Datasets.

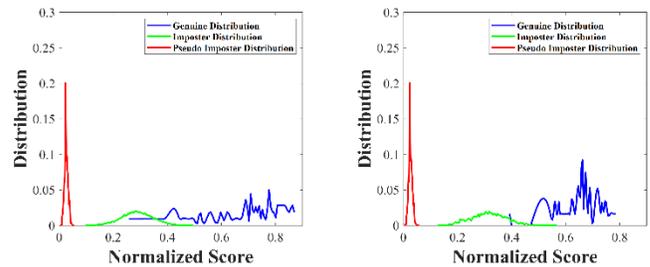


Fig. 9. Revocability Analysis using Fused Discriminant Local Hybrid Binary Gradient Contour (FD-LHBGC) Features on (a) SDUMLA-HMT and (b) UTFVP Datasets.

3) *Validations*: Obtained results from the proposed authentication system are validated by comparing with that of other existing finger vein-based authentication systems. Validation is shown on comparisons to some of the similar existing finger vein systems based on their verification performance. Table III includes similar existing verification systems implemented with variety of feature extraction and classification methods. These systems contain methods utilizing single as well as multiple instances of finger veins. Performance comparison is shown on the viewpoints of authentication performance, number of evidence/instances used for authentication and template protection.

It has been observed from Table III that the proposed verification framework outperforms the shown similar work on the combined scale of verification performance and template protection. As some of the verification systems like [5] and [8] shows better verification but at the cost of unprotected templates. Moreover, the proposed system implemented the

verification utilizing two finger vein instances contributing to a greater number of evidence as identity proofs.

The verification performance combining with template protection property from Table III indicates that the proposed-I framework i.e., with Fused-Discriminant Uniform Local Binary Pattern (FD-ULBP) features performs better than Fused-Discriminant Local Histogram Binary Gradient Contour (FD-LHBGC) features as an overall authentication system with protected templates. Finger vein feature extraction and matching time contributes to the computation cost of the overall authentication system.

Moreover, if the system includes template protection and/or utilizing more than one instance of finger vein that again adds on the computational burden. Table IV shows the comparison of feature extraction and matching time of the proposed system with some of the similar popular methods practiced for finger vein authentication tested on a varied quality finger vein dataset, SDUMLA-HMT.

TABLE III. COMPARISON OF THE VERIFICATION PERFORMANCE OF THE PROPOSED SYSTEM WITH SIMILAR EXISTING VERIFICATION SYSTEMS

Related work	Feature extraction method	Classifier	Template protection method	Verification performance (% Accuracy)		No. of instances used	Template protection
				SDUMLA-HMT	UTFVP		
Ton et al. [15]	Maximum Curvature	Correlation	---		0.4	1	No
Kauba et al. [17]	Different feature level fusion	Correlation	---		0.19	1	No
Yang et al. [18]	Anatomy Structure Analysis based Vein Extraction (ASAVE)	Elastic matching	---	1.39		1	No
Yang et al. [14]	Local Binary Pattern (LBP), t-norm based score fusion	Hamming distance	---	1.58		2	No
Ong et al. [5]	Local Histogram Binary Gradient Contour (LHBGC)	Support Vector Machine	---	0.034		2	No
Syarif et al. [8]	Enhanced Maximum Curvature, Histogram of Oriented Gradients	Support Vector Machine	---	0.14		1	No
Yang et al. [30]	Gabor Filter, Principal Component Analysis (PCA)	L2-Norm	Random proj. transform, fuzzy commitment scheme	3		1	Yes
Yang et al. [36]	Principal Curvature	Cosine similarity	Wrapping in feature domain		0.71	1	Yes
	Repeated Line Tracking	Collision Probability	Align. Robust Hash, Index of Max hash.		3.89	1	Yes
Kirchgasser et al. [35]	Gabor Filter, Linear Discriminant Analysis (LDA)	Multi-Layer Extreme Learning Machine	Biohashing, Binary Decision Diagram	7.04		1	Yes
Proposed-I	FD-ULBP	Collision Probability	GRP-IoM	0.53 ± 0.161	0.00074 ± 0.00119	2	Yes
Proposed-II	FD-LHBGC	Collision Probability	GRP-IoM	4.2 ± 0.19	1.54 ± 0.11	2	Yes

TABLE IV. COMPARISON OF COMPUTATION TIME (SECONDS) OF THE PROPOSED SYSTEM WITH SOME STATE-OF-THE-ART FINGER VEIN-BASED AUTHENTICATION SYSTEMS TESTED ON SDUMLA-HMT DATASET

Related work	Feature extraction method	Classifier	Template protection method	Computation time			No. of instances used	Template protection
				Feature extraction time	Matching time	Total Computation time		
Miura et al. [3]	Repeated Line Tracking	Miura match	---	19.24	0.97	20.21	1	No
Miura et al. [6]	Maximum Curvature	Miura match	---	0.7	0.97	1.67	1	No
Huang et al. [7]	Wide Line Detector	Miura match	---	0.56	0.97	1.53	1	No
Syarif et al. [8]	Maximum Curvature, Histogram of Oriented Gradient	Support Vector Machine	---	0.72	0.07	0.79	1	No
	Enhanced Maximum Curvature, Histogram of Oriented Gradient	Support Vector Machine	---	0.59	0.07	0.66	1	No
Ong et al. [5]	Local Histogram Binary Gradient Contour	Support Vector Machine	---	0.4496	0.0047	0.4543	2	No
Ong et al. [9]	Minutiae	GA, k-modified Hausdorff dist.	---	---	---	0.7528	2	No
Proposed-I	FD-ULBP	Collision Probability	GRP-IoM	0.4609	6×10^{-4}	0.4615	2	Yes
Proposed-II	FD-LHBGC	Collision Probability	GRP-IoM	0.4545	6.23×10^{-4}	0.4551	2	Yes

It is observed from Table IV that both the proposed multi-instance finger vein authentications are outperforming with respect to some single or multi-instance state-of-the-art finger vein-based systems on the scale of computation time. In [5], LHBGC feature extraction time is calculated with the same parameter values (number of rows and columns) as that of the LHBGC based proposed-II system for fair comparison. Feature extraction time for the proposed methods involve pre-processing time, feature extraction time, fusion time, and template protection-based code generation time for two instances of finger vein images. Despite inclusion of template protection scheme, the proposed systems are showing significantly low values for feature extraction or template generation time. Moreover, matching time for the coded template is substantially less because of simple collision computations involved. Hence, the computational complexity of both the proposed protected systems is comparatively low with respect to the shown non-protected template-based finger vein systems.

Thus, the proposed methods of multi-instance finger vein systems provide considerable authentication accuracy with lower computation cost. The proposed frameworks also facilitate renewable templates in case of compromise with high irreversibility and unlinkability to produce template protection enabled authentication.

V. CONCLUSION

The proposed multi-instance finger vein-based biometric authentication system offers significant authentication performance. The proposed system used two local texture-based features which are computationally economical. Template protection is incorporated via. highly non-invertible and unlinkable, transform-based projection offering cancelable biometric templates. Proposed framework provides significant reduction in computational cost for feature extraction and

template matching, balanced with considerably outperforming authentication accuracy. Cancelable biometric template generation method as Gaussian Random Projection based Index-of-Max (GRP-IoM) is incorporated for template protection. The proposed Fused Discriminant-Uniform Local Binary Pattern (FD-ULBP) and Fused Discriminant-Local Hybrid Binary Gradient Contour (FD-LHBGC) feature based finger vein systems are observed to outperform some existing systems on the scale of verification performance. The proposed frameworks are experimented on two standard databases as SDUMLA-HMT and UTFVP. Moreover, FD-ULBP features are found to provide more significant results than FD-LHBGC for authentication and template protection.

ACKNOWLEDGMENT

The authors would like to express their sincere gratitude towards the authorities of K.J. Somaiya College of Engineering, Mumbai for providing continual support during this research work.

REFERENCES

- [1] P. Campisi, Ed., Security and Privacy in Biometrics. Springer, 2013.
- [2] Z. Jin, J. Y. Hwang, Y. L. Lai, S. Kim and A. B. J. Teoh, "Ranking-Based Locality Sensitive Hashing-Enabled Cancelable Biometrics: Index Of Max Hashing," IEEE Transaction on Information Forensics and Security, 13 (2), 2018.
- [3] N. Miura, A. Nagasaka and T. Miyatake, Feature Extraction of Finger-vein Pattern based on Repeated Line Tracking and its application to Personal Identification, Machine Vision and Applications, Springer, 15, 2004, pp. 194-203.
- [4] T. Ojala, M. Pietikainen, and T. Maenpaa, Multiresolution Gray-scale and Rotation Invariant Texture Classification with Local Binary Patterns, IEEE Transactions on Pattern Analysis and Machine Intelligence, 24 (7), 2002, pp. 971-987.
- [5] T. S. Ong, A. William, C. Tee and M. K. O. Goh, Robust Hybrid Descriptors for Multi-instance Finger Vein Recognition, Multimedia Tools Appl., Springer Science, 77, 2018, pp. 29163-29191.

- [6] N. Miura and A. Nagasaka, Extraction of Finger-vein Patterns using Maximum Curvature Points in Image Profiles, IAPR Conference on Machine Vision applications, Tsukuba Science City, Japan, pp. 347-350, 2005.
- [7] B. Huang, Y. Dai, R. Li, D. Tang and W. Li, Finger-Vein Authentication Based on Wide Line Detector and Pattern Normalization, IEEE 20th International Conference on Pattern Recognition, Istanbul, Turkey, pp. 1269-1272, 2010.
- [8] M. A. Syarif, T. S. Ong, A. B. J. Teoh and C. Tee, Enhanced Maximum Curvature Descriptors for Finger Vein Verification, Multimedia Tools Appl., Springer Science, 76, 2016, pp. 6859-6887.
- [9] T. S. Ong, J. H. Teng, K. S. Muthu and A. B. J. Teoh, Multi-instance Finger Vein Recognition Using Minutiae Matching, IEEE 6th International Congress on Image and Signal Processing, Hangzhou, China, , 3, pp. 1730-1735, 2013.
- [10] Y. Wang, Y. Fan and W. Liao, Hand Vein Recognition Based on Multiple Keypoints Sets. In: 5th IAPR International Conference of Biometrics ICB, New Delhi, India, pp. 367-371, 2012.
- [11] Y. C. Cheng, H. Chen and B. C. Cheng, Special Point Representations for Reducing Data Space Requirements of Finger Vein Recognition Applications, Multimedia Tools Appl., Springer Science, 76 , 2017, pp. 11251-11271.
- [12] T. Ojala, M. Pietikäinen and D. Harwood, A comparative study of texture measures with classification based on feature distributions. Pattern Recognition, 29, 1996, pp. 51-59.
- [13] X. Xi, G. Yang, Y. Yin and X. Meng, Finger Vein Recognition with Personalized Feature Selection, Sensors, 13 (9), 2013, pp. 11243-11259.
- [14] Y. Yang, G. Yang and S. Wang, Finger Vein Recognition based on Multi-instance, International Journal of Digital Content Technology and its Applications, 6 (11), 2012, pp. 86-94.
- [15] B. T. Ton and R. N. J. Veldhuis, A high quality finger vascular pattern dataset collected using a custom designed capturing device, In: Proc. Int. Conf. Biometrics (ICB), Madrid, Spain, pp. 1-5, 2013.
- [16] H. T. Van, T. T. Thai, and T. H. Le, Robust finger vein identification base on discriminant orientation feature, In: 7th Int. Conf. on Knowledge and Systems Engineering (KSE), Ho Chi Minh City, Vietnam, pp. 348-353, 2015.
- [17] C. Kauba, E. Piciuccio, E. Maiorana, P. Campisi, and A. Uhl, Advanced variants of feature level fusion for finger vein recognition, In: 2016 International Conference of the Biometrics Special Interest Group (BIOSIG), Darmstadt, Germany, 2016.
- [18] S. Qiu, Y. Liu, Y. Zhou, J. Huang, and Y. Nie, Finger-vein recognition based on dual-sliding window localization and pseudo-elliptical transformer, Expert Systems with Applications, 64, 2016, pp. 618 - 632.
- [19] L. Yang, G. Yang, Y. Yin, and X. Xi, Finger vein recognition with anatomy structure analysis, IEEE Transactions on Circuits and Systems for Video Technology, 28 (8), 2018, pp. 1892-1905.
- [20] A. Banerjee, S. Basu, S. Basu, and M. Nasipuri, Artem: a new system for human authentication using finger vein images, Multimedia Tools and Applications, 77, 2018, pp. 5857-5884.
- [21] H. Hong, M. Lee, and K. Park, Convolutional neural network-based finger-vein recognition using NIR image sensors, Sensors, 17 (6) , 2017, 1297.
- [22] R. Das, E. Piciuccio, E. Maiorana, and P. Campisi, Convolutional neural network for finger-vein-based biometric identification, IEEE Trans. Inf. Forensics Secur., 14 (2), 2019, pp. 360-373.
- [23] G. K. Sidiropoulos, P. Kiratsa, P. Chatzipetrou and G. A. Papakostas, Feature Extraction for Finger-Vein-Based Identity Recognition, Journal of Imaging 7 (5), 2021, 89.
- [24] M. Gomez-Barrero, C. Rathgeb, J. Galbally, J. Fierrez, and C. Busch, Protected Facial Biometric Templates Based on Local Gabor Patterns and Adaptive Bloom Filters, in Proc. Int. Conf. Pattern Recognit., Stockholm, Sweden, pp. 4483-4488, 2014.
- [25] G. Li, B. Yang, C. Rathgeb, and C. Busch, Towards Generating Protected Fingerprint Templates Based on Bloom Filters, in Proc. Int. Workshop Biometrics Forensics (IWBF), Gjovik, Norway, pp. 1-6, 2015.
- [26] J. Bringer, C. Morel, and C. Rathgeb, Security Analysis of Bloom Filter-based Iris Biometric Template Protection, in Proc. Int. Conf. Biometrics (ICB), Phuket, Thailand, pp. 527-534, 2015.
- [27] P. P. Paul and M. Gavrilova, Multimodal Cancelable Biometrics, IEEE 11th International Conference on Cognitive Informatics & Cognitive Computing, Kyoto, Japan, pp. 43-49, 2012.
- [28] Y. Chin, T. Ong, A. Teoh and K. Goh, Integrated Biometrics Template Protection Technique Based on Fingerprint and Palmprint Feature-level Fusion, Information Fusion, 18, 2013, pp. 161-174.
- [29] W. Yang, J. Hu, and S. Wang. A finger-vein based cancellable biocryptosystem. In: Network and System Security 7th International Conference, NSS 2013, Madrid, Spain, pp. 784-790, 2013.
- [30] W. Yang, S. Wang, J. Hu, G. Zheng, J. Chaudhry, E. Adi and C. Valli, Securing Mobile Healthcare Data: A Smart Card Based Cancelable Finger-vein Bio-Cryptosystem, IEEE Access, Special Section on Cyber Threats and Countermeasures in the Healthcare Sector, 6, 2018, pp. 36939-36947.
- [31] Y. Liu, J. Ling, Z. Liu, J. Shen, and C. Gao. Finger vein secure biometric template generation based on deep learning. Soft Comput., 22 (7), 2018, pp. 2257-2265.
- [32] M. Gomez-Barrero, C. Rathgeb, G. Li, R. Ramachandra, J. Galbally, and C. Busch. Multi-biometric template protection based on bloom filters. Information Fusion, 42, 2018, pp. 37-50.
- [33] D. Hartung, M. Tistarelli, and C. Busch, Vein minutia cylinder-codes (VMCC), In: International Conference on Biometrics, ICB 2013, Madrid, Spain, pp. 1-7, 2013.
- [34] T. Ong, A. William, T. Connie and M. Kah Ong Goh, "Robust hybrid descriptors for multi-instance finger vein recognition", Multimedia Tools and Applications, 77(21), 2018, pp. 29163-29191.
- [35] S. Kirchgasser, C. Kauba, Y. Lai, J. Zhe and A. Uhl, Finger Vein Template Protection based on Alignment-Robust Feature Description and Index-of-Maximum Hashing, IEEE Transactions on Biometrics, Behavior and Identity Science, 2020.
- [36] W. Yang , S. Wang , J. Hu , G. Zheng , J. Yang and C. Valli, Securing Deep Learning Based Edge Finger Vein Biometrics With Binary Decision Diagram, IEEE Transactions on Industrial Informatics, 15 (7), 2019, pp. 4244-4253.
- [37] H. O. Shahreza and S. Marcel, Towards Protecting and Enhancing Vascular Biometric Recognition Methods via Biohashing and Deep Neural Networks, IEEE Transactions on Biometrics, Behavior, and Identity Science, 3 (3), 2021, pp. 394-404.
- [38] B. Prommegger, C. Kauba and A. Uhl, Multi-Perspective Finger-Vein Biometrics, IEEE 9th International Conference on Biometrics Theory, Applications and Systems (BTAS), 2018, pp. 1-9.
- [39] M. Haghghat, M. A. Mottaleb and W. Aalhalabi, Discriminant Correlation Analysis: Real Time Feature Level Fusion for Multimodal Biometric Recognition, IEEE Transactions on Information Forensics and Security, 11 (9), 2016, pp. 1984-1996.
- [40] Y. Yin, L. Liu, and X. Sun, "SDUMLA-HMT: A multimodal biometric database," in Proc. Chin. Conf. Biometric Recognit, Springer Verlag, pp. 260-268, 2011. (accessed 05 Jan. 2019).
- [41] Twenty University dataset, "http://www.utwente.nl/em/eemcs/ds*", online accessed on Oct-2019.

SHD-IoV: Secure Handover Decision in IoV

Hala E. I. Jubara

Department of Information System
Faculty of Computer and Information System
Jouf University, Sakakah, Aljouf, Saudi Arabia

Abstract—Internet of Vehicle (IoV) is the smartest thing being connected over the Internet. With continuously increasing urban population and swiftly growing of cities, causes moving vehicles with various speeds. These high speeds may increase the handover delay (HoD), accordingly causing an insecure connection due to the handover interruption. For instance, some of the network protocols try to overcome the problem without considering transport layer supports. This article proposes a dynamic HO algorithm with a cross-layer architecture called Secure Handover Decision (SHD) in IoV to assist the protocol layers aware of consecutive HOs of the vehicle. The results show that vehicle communication in IoV is more secure and lossless by reducing HoD in both sides of vehicle and network during fast movement.

Keywords—Internet of Vehicle (IoV); HO; L2; Stream Control Transmission Protocol (SCTP); Secure Handover Decision (SHD); security

I. INTRODUCTION

The need for mobile internet connectivity has increased as more users travel from place to place, for example, town to town over long distances. In a vehicular network, the user's vehicle connects to the internet through a fixed infrastructure installed on the side of a road for vehicle-to-infrastructure (V2I) communication. Another type of vehicle network communication is vehicle-to-vehicle (V2V) communication, which addresses the transmission of information among vehicles. However, this type of infrastructure contains gateways as well as BSs that offer services like the Internet of Vehicles (IoV) [29, 35, 36, 37, 38, 39]. In the IoV network, the vehicles are highly dynamic and can move at a high speed, which may cause a high handover (HO) rate leading to a communication delay or disruption (Fig. 1). Additionally, V2I communication is expected to meet many difficulties like poor channel quality plus connectivity because higher vehicle speeds lead to HO delay. Thus, there is a crucial need for efficient communication such as protocol or BS type communication that considers the specific characteristics of vehicular networks [4, 16, 17, 25, 26].

From the protocol side, most of the network layer (L3) protocols have a long HO delay, which affects the communication of the vehicles while moving. On the other hand, the current transport layer protocols suggested for better mobility can't address mobility alone because most of these ideas rely on the network layer mobility management necessities for the handover. Their proposal is purely to reduce the degradation in the performance of the transport layer caused by the handover. Several of these newly evolving protocols for L4 such as mobile Stream Control Transmission

Protocol (*mSCTP*), offer a basis for mobility support because they have multi-homing features that allow a mobile station to use a new IP address, while still assigning the previous IP address [5-9, 34].

In a data link layer (L2), an HO delay compromises the BS in completing the HO procedure with the next target BS (TBS) along the vehicle's path. Many network technologies such as cellular networks (GSM, 3G, 4G standard) [1, 23, 33] have been developed for broadband wireless access to meet the demand for high data rates in the wireless service. The most important improvement in this type of network for maintaining mobility is HO support. The HO is performed to maintain a continuous data-transmission service for all applications when the user is moving across the cell borders of the BSs. Three basic types of HO [3] have been defined for cellular networks: a hard handover (HHO), the macro-diversity handover (MDHO), and fast base station switching (FBSS). MDHO and FBSS are soft optional handovers, whereas HHO is a mandatory handover in WiMAX and LTE systems. HHO adopts a break-before-make method, where the user stops its radio link to the serving BS before establishing its radio link with the target BS [30, 32]. Because HHO is a simple method, it causes a long HO delay and disrupts service for certain applications, especially when the user is traveling at a high speed, such as traveling on a highway.

In a vehicular environment, as the vehicles are moving, the traffic generated by other background vehicles connected to the same BS decreases the available amount of bandwidth as the collision rates at the data link layer increase [31]. Under these network conditions, the HO may trigger repeatedly even for a static wireless station.

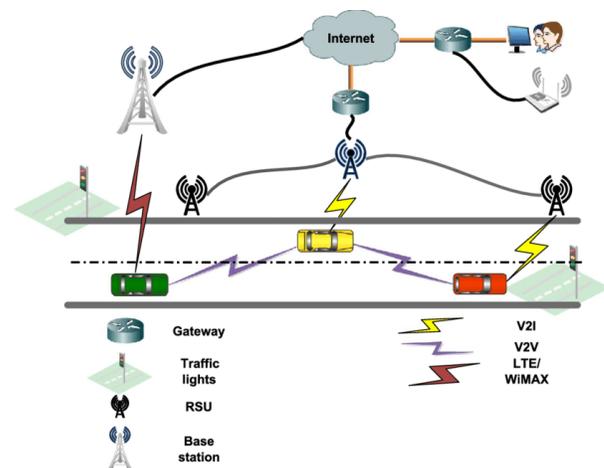


Fig. 1. IoV Networks Model.

This paper discusses a case in which a vehicle moves at a high speed from one BS to another in the IoV scenario, and it will cause long HO delays to the current Internet connection. To reduce this delay accordingly packet-loss rate during movements of the vehicle, an enhancement over the existing mSCTP protocol in L4 to support mobility has been proposed. This achieves through a cross-layer design of L2 and L4 to optimize the performance in terms of HO delay at L2, L4, and L3 consequently. The cross-layer design generates an L4 awareness regards to the vehicle movement using the radio signal strength indicator (RSSI) in L2 and utilizes the LM to track the vehicle movement with high speed along with the network.

The rest of this paper is organized as follows. An overview of the previous related studies is introduced in Section II in terms of mobility management of the protocol layers, vehicle, and cross-layer as well. The framework of vehicle mobility management to overcome the stated problem at high speed is discussed in Section III. Section IV presents the idea of the work of the smooth adaptive handover management for vehicle users. Then Section V shows the simulation test and Section VI details the results and analysis. Finally, Section VII provides some concluding remarks regarding this research.

II. RELATED WORK

A. Management of Mobility in different Protocol Layers

1) *Network layer mobility solutions:* The most common examples of network layer mobility solutions are Mobile IPv4 (MIPv4) and Mobile IPv6 (MIPv6) [4]. The Mobile IP permits transparent packet routing to the mobile user, as opposed to each node being allocated a permanent IP address that correlates to the home network. Furthermore, when a mobile user roams across several foreign subnets, each subnet receives a new IP address (Care-of-Address (CoA)). The mobile user then sends a binding update to its home agent (HA), which keeps track of the node CoA's current location and tunnels traffic from the mobile user to the mobile user. Routing optimization, hierarchical, and predictable algorithms have all been reported as breakthroughs for Mobile IP handover [5]. Furthermore, by tunneling traffic to the mobile user's AP, triangle routing is avoided [14].

The network is divided into domains by Hierarchical Mobile IPv6 (HMIPv6) [4], [8], each of which contains numerous access routers (AR) and a Mobility Anchor Point (MAP) that links the domain to the Internet. The MAP takes mobile user packets and tunnels them to the domain level CoA, as well as controlling the domain's mobility. This reduces changeover delay and loss by completing a micro-level address while still doing the macro-level handover, which has a large latency.

Fast MIPv6 (FMIPv6) [8] as a Fast Handover Protocol that uses L2 triggering for handover to improve speed and decrease packet loss. This is performed by announcing the existence of mobile users as well as the new AR's readiness to receive data from the new CoA. The necessity for collaboration between the user and for both prior and new AR, as well as the high unpredictability of packets arriving at the Aps is the system's

main drawbacks. In a comparison of several ways, FMIPv6 outperforms HMIPv6 in terms of handover delay and packet loss, however using both methods improves performance through each of them alone [4, 8, 15, 22].

2) *Transport layer mobility solutions:* the TCP and UDP protocols have been enhanced to provide mobility transport layer protocols, which are still the most commonly used on the Internet [1, 2, 11, 12, 21]. Another of these protocols is Stream Control Transmission Protocol (SCTP), which allows each endpoint of an association to utilize several IP addresses, allowing a mobile user to multi-home. The mobile Stream Control Transmission Protocol (mSCTP), which uses the SCTP IP address extension to allow an association's terminals to change their main IP address without breaking their current connection [21][24], is another innovation. Even while mSCTP can provide precise conditions for faultless handover when the main address must be changed, there is still an issue. Even while mSCTP can provide precise conditions for faultless handover when the main address must be changed, there is still an issue. A cross-layer design across several levels has been proposed in several studies [1, 2, 10] to improve the mobility of transport layer protocols such as SCTP and mSCTP. They were able to demonstrate that SCTP can give lower handover latency than mobile IP and a much reduced handover latency for several different types of handover in such studies.

3) *Cross-Layer mobility management for vehicle users:* Several solutions [11, 13, 14] that seek to promote smooth handover in high-speed users (e.g. cars) were explored. The authors of [11] utilize a system that forecasts vehicle motion to optimum performance in a high-speed environment, and they estimate that there will be no concerns as the length of connectivity increases. In the 802.21 method, the authors of [14] adopt a previous knowledge technique wherein network information is collected from both the mobile user and the network infrastructure in order to establish a connection with a new subnet ahead of time. A similar research [13] proposed lowering the effect of a service outage among high-speed users. This proposal offered a packet forwarding control which would select a point of agreement for forwarding packets in order to transmit them through a shorter path during a handover. The author in [14] proposes a network mobility protocol (NEMO) for usage in a vehicle networks (VANET) environment on a roadway. Despite the fact that each vehicle is traveling at a high rate and in a fixed direction in this case, vehicle-to-vehicle (V2V) connections might provide the vehicle with an IP address.

B. Cross-layer Mobility Solution

Various cross-layer efforts have been created in an attempt to reduce the HO delay. The author in [26] describes VSPLIT, transport layer performance improvement architecture for Internet-based Vehicle-to-Infrastructure (V2I) communications in vehicular networks based on TCP cross-layering and splitting methods. The primary goal of this strategy is to enhance TCP handover performance in 802.11 networks. The

VSPLIT-TCP cross-layer TCP protocol, which uses IEEE 802.21 Media Independent Handover (MIH) services to modify congestion control during the changeover by learning various network parameters after the handover. SHSBM, a Smooth Handover Scheme based on mSCTP, is proposed in the literature of [27]. To best support fast-moving users, SHSBM takes use of SIGMA [7, 10, 18, 19, 20, 21] and employs Buffer and Tunnel. They also provide two ways for dealing with the issue presented by the Buffer-scheme—sequence Out of Order. In comparison to SIGMA and Mobile IPv6 upgrades, performance criteria such as packet loss rate, throughput, and handover time were used to evaluate performance.

In their study, [28] provides a framework for linking vehicle networks to the IPv6-based Internet. This concept provides a road domain-based architecture to minimize the frequency of mobility handovers. In this study, they are developing a distributed address configuration mechanism for car networks. Using this method, a vehicle obtains a unique address from the nearest access point (AP), avoiding the detection of duplicate addresses. On the basis of this architecture, a routing mechanism based on geographical position is suggested. A car connects to the Internet by connecting to the nearest access point, and the routing algorithm has been applied to the link layer. During the mobility procedure, the vehicle's home address is always used to identify it, and no care-of address is necessary. As a result, packet loss due to a change in address is avoided. Additionally, packet loss is greatly reduced since a vehicle can receive data from the same AP during the mobility changeover phase. Their approach can minimize communication latency and packet loss, but IPv6 introduces a new delay that can affect upper-layer connectivity. They offer a cross-layer rapid handover strategy that communicates physical layer information with the link layer to decrease handover delays in automobile networks. The WiMAX mobile multi-hop relay mechanism, which allows inter-vehicle communications to connect to the Internet through a relay vehicle, provides the foundation for this technique. However, IP mobility is not included in the program. The need for flawless communication in high-speed settings is an appealing and difficult problem that necessitates accurate IoV in most modern networks [35, 36, 37]. While the majority of the preceding work focused on changeover for and moderate speeds, the requirement for smooth communication under high-speed situations is an appealing and demanding issue since most new networks require precisely IoV. In this situation, employing the lower layer's handovers and the transportation layer's communication layer will make handover awareness and avoid communication interruption, minimizing packet loss and, as a result, increasing network QoS.

III. VEHICLE MOBILITY MANAGEMENT FRAMEWORK

A. System Architecture

Any moving user traveling at varied speeds while communicating over the same network technology can use the SHD architecture. The vehicle and BS modules are the two most important modules in the design. A graphical representation of this is shown in Fig. 2. The vehicle module is in charge of protocol design and handles one SCTP relationship with a Domain Name Server (DNS) entity.

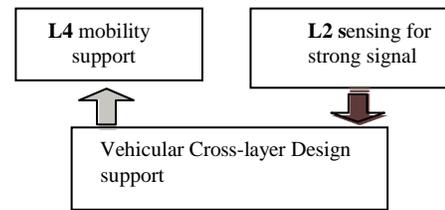


Fig. 2. Mobility Management's System Architecture.

The DNS entity, on the other hand, monitors and tracks the vehicle's mobility via the Dynamic Host Configuration Protocol (DHCP) server [6-9] (as shown in Fig. 5 and Fig. 7), which first monitors and tracks the vehicle's global movement by saving the current vehicle address in the server across networks to support L4 multi-homing. Second, the BS module controls the vehicle's HO using an adaptive algorithm in L2 that is dependent on factors like the received Signal to Noise Ratio (SNR), Received Signal Strength Indicator (RSSI), and vehicle speeds. The flow under this design begins when the vehicle initiates a handover to the TBS at a specified speed. Once the BS module has assessed speed and RSS/SNR, HO signals are delivered to the vehicle through the network. When the vehicle module gets the information marking the start of HO, cross-layer communications are sent. As a result, the SHD approach updates upper layers in response to the rapid change in speed.

B. Vehicle Mobility Management using Cross-Layer Design

The cross-layer design is suggested in this paper for managing transport layer mobility. The design proposes the mSCTP transport layer protocol, which is based on transport signaling messages and supports vehicle mobility handover over an IP network. As a result, the SHD's purpose is to ensure that any protocol may be used at any tier. In this case, the SHD's HO method permits information about the HO choice to be shared between layers. The next sections go over HOs at the L2, L3, and L4 levels in great depth.

1) *L2 HO Delay*: The WiMAX BS delay is employed to accomplish the HO operation at the datalink layer in this study. Signal strength is routinely measured in these types of BSs using parameters like the Received Signal Strength Indicator (RSSI) [34]. As a result, the HO is started as soon as the RSSI from the presently serving BS falls below a certain level. When the HO is necessary, this threshold is fixed (2dBm for traditional WiMAX) and is utilized to launch and execute it. When communication quality deteriorates, the vehicle's L2 looks for the best BS for HO and uses it as a TBS. We reduce the time here by restricting the number of scanning TBS to three (Fig.3, N=3).

The impact of an L2 delay on SHD performance may be split into two categories. The influence of the BS's HO process time is the first, while the vehicle speed is the second. L2 initiates a handover and the scanning procedure for the TBS begins if the SBS signal quality deteriorates, which takes roughly 15 ms for a high-speed vehicle [24]. Until the TBS completes the HO, communication is disabled. The whole

delay of L2 compromises the synchronization time (T_{sync}) between BSs and frame duration:

$$T_{L2} = T_{sync} + T_{frame} \quad (1)$$

Upon synchronization with the arriving downlink for other HO messages related to the BS HO, the downlink packet may be broadcast immediately (DAD procedure, tunneled packets, delay of each hop in a wired, resolution procedure, ranging process, re-authorization during HO, and re-registration). The L3 HO delay explains the role of L3 at the HO delay time.

2) *L3 HO Delay*: The network layer delay of handover is roughly 1 minute due to DAD and other HO messages on the network. L3 HO delay can be sent over a cross-layer of L2 and L4, allowing the delay to be linked solely to L2 and L4 HO. The issue is that the SCTP relies on the LM/DNS to keep track of the vehicle's current location when the IP domain changes. As shown in Fig. 4, it may be done during SCTP's HO, when the vehicle L4 updates the LM with the updated BS after HO. This enables real-time tracking of the new car. The performance of the L2 HO at various vehicle speeds, as well as the computations required for a successful HO to adjust L4 to each speed are discussed in the next section.

3) *L4 HO Delay*: The protocol detecting a HO causes L4 delay, which might last several seconds depending on the round-trip duration (approximately 10 ms) between the vehicle and the CN. This might lead to packet loss and, as a result, a decrease in throughput. The mSCTP, on the other hand, is utilized to facilitate multi-homing when going in a fast vehicle. The vehicle HO process is depicted in Fig. 3 from the time the vehicle gets the network's HO decision (SBS) until the time the dynamic HO is executed at L2.

To complete the HO between the vehicle and CN, the ASCONF SET PRIMARY/DEL IP messages that cause the HO delay at L4 are necessary. Because the connection latency for updating the LM has no influence on the SHD handover delay, the time necessary to update the LM of REG.REQ/RSP is disregarded. As a result, the L4 transfer's total interruption time is:

$$T_{L4} = T_{(ASCONF SET-PRIMARY/DEL-IP)} + RTT \quad (2)$$

4) *Adaptation between L2 and L4*: The vehicle adapts the L4 protocol SCTP and the vehicle speed at L2 using algorithms. At varying speeds, this technique dynamically manages the SCTP protocol's handover decision. It runs the vehicle's L2 protocol in order to make a HO decision based on the SBS's current signal quality, which is indicated by the RSSI in the MOB-NBR-ADV message. On the other hand, depending on this number, the HO execution produces the strongest TBS signal. Fig. 3 shows the flow of the HO algorithm. As shown in this picture, when a vehicle enters the HO region of the TBS, it receives a message about the availability and amount of TBSs. The algorithm leverages the vehicle speed supplied by the BS to make a dynamic HO choice when the vehicle signal strength begins to decline. To

relate the vehicle's speed to the HO choice, the computer use Equation 3.

$$Th_{HO} = Th_{loss} (1 + \log_2 (v+1)) \quad (3)$$

Furthermore, the adaptive method is based on the following conditions to avoid performing unnecessary actions such as lengthy HO delays or squandering network resources with unnecessary HOs, both of which can result in substantial system performance degradation:

$$RSSI_{SBS} < TH_{HO} \quad (4)$$

$$RSSI_{TBS} > Th_{loss} + \Delta D \quad (5)$$

When the RSSI falls below the TH_{HO} in Eq.4, the HO operation will start. In addition, in Eq.5, the HO is only done if another BS has an RSSI that is at least D greater than the Th_{loss} . These equations change the handover threshold (TH_{HO}) based on the current vehicle speed (v) and RSSI of the SBS, which have varying values at different points in the coverage area. To make the threshold dynamically adapt with speed, Eq.3 mentions the link between the TH_{HO} and the speed v . The communication's loss threshold (TH_{loss}) and the hysteresis value D govern the TBS. When the vehicle's speed rises, TH_{HO} rises as well, and the vehicle executes a straight handover to the following TBS to prevent a delay. TH_{HO} , on the other hand, uses Th_{loss} to achieve the lower limit when the speed is low. After that, the system compares TH_{HO} to the current RSSI (as in Eq.4 and Eq.5). The selection is made based on the vehicle's speed as well as the RSSI. As a result, two scenarios are examined for a HO operation. If TH_{HO} is bigger than RSSI (as in Eq.4), the HO operation is first carried out at the highest TBS signal intensity. Otherwise, the BS executes the HO by comparing the SNR with the neighboring BSs (NBSs). The second option is taken to minimize communication interruption due to fast changes in the received signal level caused by distortion or short-term shadowing of high-speed vehicles (Eq.5).

Due to the numerous HOs that occur at greater speeds, the adaptive HO algorithm's objective is to avoid a delay and packet loss during transmission. TH_{HO} and Th_{loss} are computed for each TBS at each handover (the method for initiating and performing a HO from the SBS to the TBS of vehicular users) (Eq.3). Because the HO delay is small, packet loss does not need a drop in the packet loss rate if the adaptive algorithm effectively controls the occurrence of a HO. Due to the additional latency, our approach employs an upgraded SCTP to decrease packet loss.

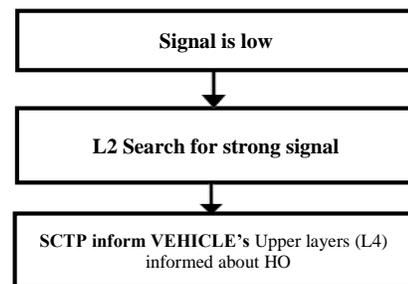


Fig. 3. Proposed Speed-Adaptive Algorithm.

IV. SECURE HANDOVER DECISION FOR VEHICLE USER

To assure the system's simplicity, adaptability, and efficiency while also achieving various aims, a cross-layer design of a SHD was carried out. Our solution, on the other hand, may be used to minimize HO in any protocol layer by changing user settings. As demonstrated in Fig. 4, the idea of a SHD-SCTP is that information may be transmitted across the vehicle's many protocol levels using primitives (short messages between layers) at L2, L3, and L4. A cross-layer design can help with mobility management by reducing HO delays and improving performance.

A. The Proposed Secure Handover Decision (SHD)

Each layer offers the higher layer with encapsulated services to use the information in that layer, focusing primarily on the L2 and L4 information exchanges (as in ISO protocol levels). This data is used to modify the L4 protocol architecture to changes in vehicle speed as follows:

The car is traveling at a rapid rate from the SBS to the TBS, and the signal strength of the SBS is deteriorating at this moment, resulting in communication deterioration. The car then enters the TBS through the handover area, and the TBS' signal strength begins to grow. L2 transmits a LinkStatusChange.end message to the upper layer network layer at this point (L3). When the vehicle arrives at the handover location and the connection with the SBS is lost, L2 uses LinkConnect.ind to send a message to L3 requesting the available number of TBSs. L2 has received a LinkUp.ind, indicating that the signal strength is growing, and a message from L3 alerting L2 that the network has been reached in the last phase of the handover, which is the conclusion of the handover. The flow of messages at the user side during handover is depicted in Fig. 4, which is a flow chart of the cross-layer design.

The L2 connections/disconnections are synchronized with the mSCTP flow thanks to the cross-layer SHD architecture. The TH_{HO} of active senders is set to a value that is determined by the vehicle's current speed and the TBS's updated RSSI. This is done right before the handover, when the car is removed from the SBS and no BS or mSCTP handover is required. BS signaling is used to get this information from the vehicle's BS. This improved handover decision can help real-time applications prevent packet loss or significant delays, while also increasing network efficiency and user fairness.

The mSCTP communications are unfair because to the various speeds of the vehicle nodes. Because quicker users have a larger number of handovers in the same amount of time as slower users, they often receive fewer throughputs. Furthermore, the standard TH_{HO} requires some time to reach the right functioning point before the handover, which takes longer when additional (slower) users are present in the HO region between two BSs. When a new connection to or disconnection from the BS occurs, the TH_{HO} can establish the correct HO choices for the SCTP flows. Because an SDH does not implement any L3 protocols to minimize the HO latency, this allows for a reduction in the disparity between fast and sluggish nodes.

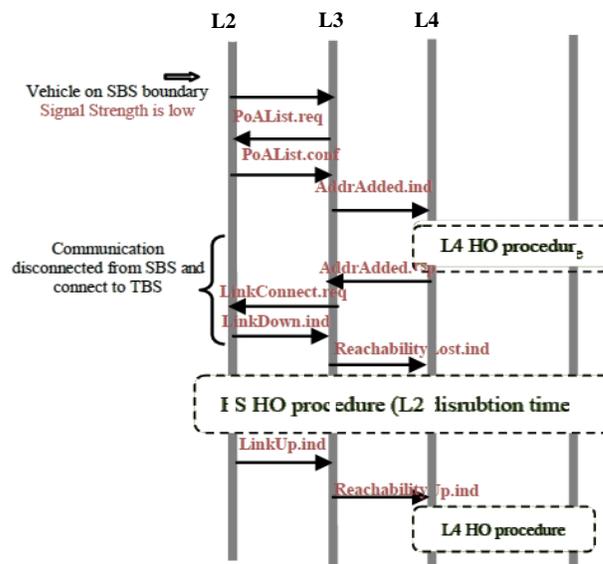


Fig. 4. Secure Handover Design.

B. Handover Procedure

A timing diagram depicting the cross-layer design is presented in Fig. 5. This design includes the two protocol levels' handover procedures (L2 and L4) as well as the cross-layer design's delay. The handover delay in L2 involves BS signaling messages between the SBS and the vehicle to begin (trigger) and conduct a typical HO procedure. The following communications come from the vehicle's L2 to the top levels, instructing them to begin the HO in L4.

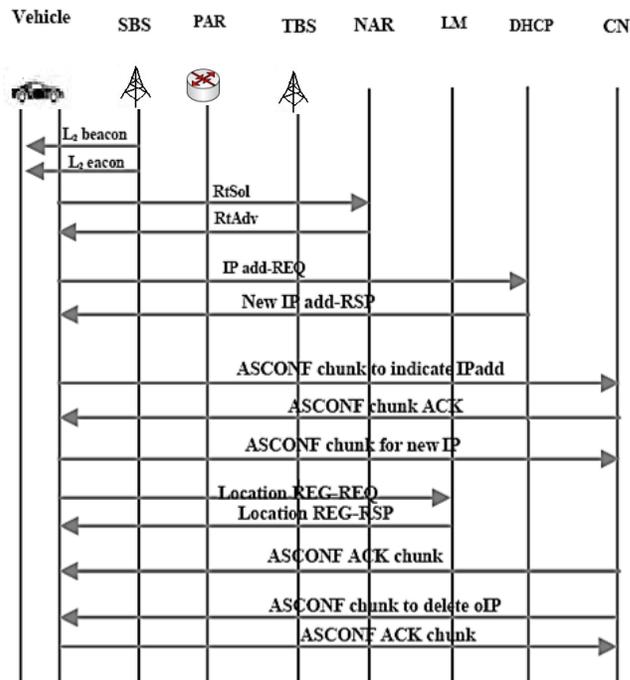


Fig. 5. Timing Diagram of the Proposed Idea Registration Request and Response (REG.REQ/RSP) is Minimal [20] and it is the Final Message in the HO Stages.

However, the majority of the L4 HO delay in this architecture is due to SCTP's set primary chunk as well as removing old IP (ASCONF SETPRIMARY/DEL-IP) handover messages, as well as the RTT of messages between the vehicle and CN (about 1–10 ms). This HO process is a conventional SCTP procedure with the addition of the LM/DNS server in the design. Because the handover delay is unaffected by the connection delay in updating the LM, the time for the location.

Finally, as shown in Fig. 5, the delay of our cross-layer design between L2 and L4 of the vehicle is around 34s, which is insignificant when compared to the L2 delay time. A cross-layer design's overall handover latency may be computed as follows:

$$(T_{HO}) = T_{L2} + T_{L4} (\text{ASCONF SET-PRIMARY/DEL-IP}) + \text{RTT} \quad (6)$$

The data connection layer delay is TL2, while the transport layer delay is TL4. We can eliminate the L3 duplicate address detection (DAD) delay, which is connected to the new address through the LM, and update the vehicle position at the TBS without an additional delay using this architecture. The LM can also be used to solve the problem of triangular packet routing between the CN and the vehicle. Because the CN continually delivers packets to the vehicle's current address across the LM, the mSCTP may work collaboratively with the LM to decrease the handover latency along with different layers. The interruption time from L2 is around 10 ms, which is minimal for L3. The HoD for this design is estimated from the vehicle to the CN. ASCONF to SET-PRIMARY/DEL-IP takes around 0.045 milliseconds in the L4 protocol, resulting in a total handover latency of about 20 milliseconds.

C. Design Goals

The following are the key objectives of this design:

1) As a mobile node, a vehicle must be connected to the network internationally. The SDH approach, on the other hand, accomplishes this purpose by employing a DNS server and an LM to track the vehicle's present location and forward packets quickly.

2) The whole vehicular network is utilized. This is a good goal for increasing mSCTP performance on the IoV network, since the protocol suffers from a large number of handovers. Our goal is to maximize the throughput of the SCTP flows before any losses or other delays occur. Between conflicting speeds and mSCTP flows, a fair handover choice is made. Handovers conducted by vehicle users traveling at various speeds might result in unfair behavior in the mSCTP. Users that stay connected to the same BS for a long time obtain better throughput in present mSCTP implementations because they experience fewer handovers. Furthermore, users who drive at fast speeds do not have enough time to receive a HO at the proper operating point. By swiftly tailoring T_{HO} to the vehicle speed and network circumstances (i.e., SNR), our handover technique can decrease changeover latency and interruption time, ensuring improved fairness between different vehicle speeds and competing SCTP flows.

V. SIMULATION TEST

A. Simulation Environment

The simulated architecture illustrates that the vehicle is traveling at high speeds along the highway (70–120 km/h) and is connected to the network through the IoV (Fig. 6). The coverage area of each BS that links automobiles to the Internet is about (1000 - 10000 m), with a 200-meter overlap between the two BSs. On the network side, the BSs are connected through the AR, with every two ARs connected to one MAP. This scenario creates an IoV communication by joining the network directly. The upper component of the network, as illustrated in the diagram, links the vehicle's present position and transmits traffic to it according to IoV services. The OMNET++ simulation was utilized to assess this architecture, together with MATLAB to compare network settings.

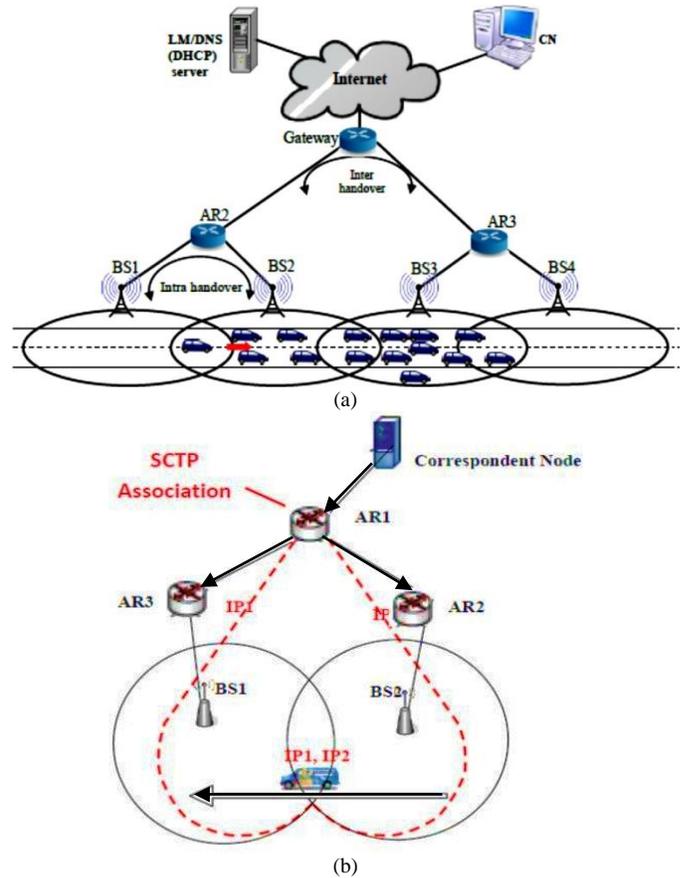


Fig. 6. Network Scenario. (a) One Vehicle Scenario. (b) Background Traffic Scenario.

The simulation performance compares two simulation models in the following way:

1) Scenario A: a single vehicle mobility management system (Fig. 6(a)). In the single vehicle instance, a vehicle drives in a straight path from the SBS to the TBS zone, with no traffic or network load. That is, the background traffic will have no effect on the car. In this instance, the handover may be limited to simply one vehicle.

2) Scenario B: Traffic mobility management in the background (Fig. 6(b)). The car is going ahead to the TBS with background traffic in the Background traffic mobility example. The HO happens when there are ten cars on the network, indicating that the vehicle is affected by the crowded network. In this situation, the car would be handed over after a longer period of time.

In order to detect and assess the performance of the proposed design and its impacts on the background traffic, we measured the HO latency and throughput for data transfer as a performance parameter of the system.

B. Background Traffic Implementation

Ten automobiles are deployed as background traffic inside the coverage area of the TBS on the network in this simulated scenario to assess the network's performance. Every vehicle travels at a different pace, and multiple of them communicate with their own networks, causing network congestion. This background traffic is created in two phases, each of which correlates to a different car count (up to ten). Each stage has different traffic levels, such as one car in the first and 10 vehicles in the second.

The background traffic conveyed to the SBS by other cars (Fig. 7(b)) raises the loss rates at L2 and hence limits the amount of available bandwidth. Even for static moving vehicles, this is a crucial element that impacts load variation and, as a result, activates the HO. The job of HO control in this situation is combined with the load-balancing service necessary to maintain an optimal decision point for deciding HO. This design evaluates performance in a variety of scenarios, such as background traffic.

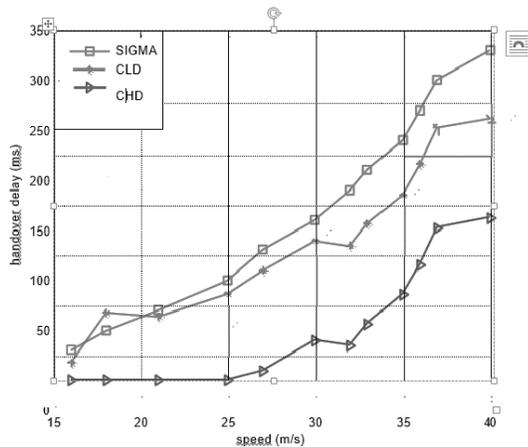


Fig. 7. HO Delay Comparison.

VI. RESULTS AND ANALYSIS

A. Performance Evaluation

To assess this concept, SHD compares three mobility options in terms of scalability, as measured by the number of vehicles executing simultaneous handovers and vehicle speed. For scalability, two mobility scenarios were investigated, in which a single vehicle and ten vehicles, respectively, transit the overlapping region at varied speeds between 10 and 40 m/s. As demonstrated in the findings, this simulation can assess the

ability of each handover strategy to maintain a shorter HO latency in various vehicle mobility models with changing network characteristics. The following sections go over the performance in further depth.

1) *Handover evaluation:* First, when the triggering time of L2 is roughly 15ms for the BS, the total HO latency of the SHD design is compared to the other design advancements. This L2 HO latency is consistent with what is seen in networks for high-speed users. When the traveling speed is increased to 40 m/s, as shown in Fig. 8, the HO delay of SHD is clearly reduced. This is because while the car communicates with the CN via the old way, it may simultaneously do L2 triggering on the other user interface. As a result, as compared to the other design advancements, the impact of these latencies can be significantly reduced (SIGMA). Because there is insufficient time for a vehicle to prepare for a new course, the HO delay of the SIGMA upgrades is roughly 2.40–2.49 s, which is substantially greater than that of the SHD design.

The HoD between vehicles is around 20ms, depending on the RTT to CN. Fig. 8 and 9 illustrate a comparison between the proposed design and existing HoD designs, while Fig. 10 displays the HoD when the network load is high. Four different scenarios were evaluated to validate the concept, as illustrated in Fig. 9.

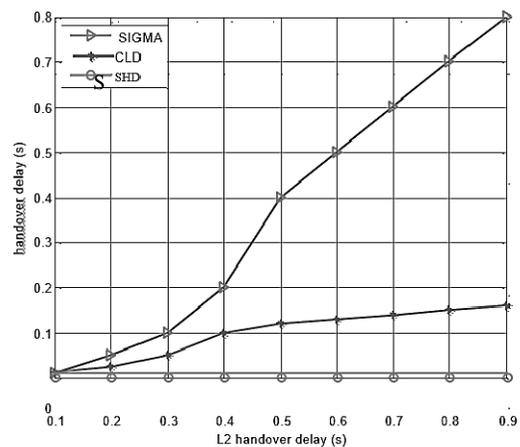


Fig. 8. Impact of an L2 HO Delay.

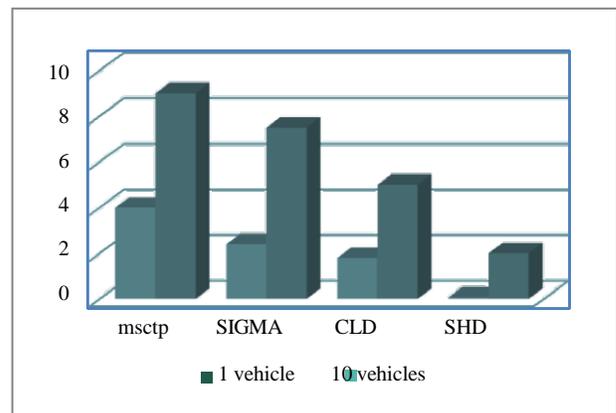


Fig. 9. Handover Delay with Background Traffic.

The first design employs the mSCTP to support a HO during a speed fluctuation; this design uses a cross-layer design to update L4 with current speed [1]. The second design (SIGMA) employs IP diversity in conjunction with SCTP to provide a multi-homing HO mechanism through the LM without the usage of L2 or a cross-layer [5-7]. The third design, SHD, is a cross-layer design between L2 and L4 enabling a speed-independent handover. However, the most recent design SHD uses an adaptive algorithm to create an ideal seamless HO during high-speed vehicle movement using a cooperative cross-layer mechanism between L2 and L4. The numbers show the outcomes of the tests.

2) *Throughput and packet loss*: For different vehicle speeds, communication time in one BS coverage region is around 67s, and HoD is about 25ms. This indicates that for high-speed automobiles with a repeating HO, the vehicle is unable to receive packets for 0.2 seconds before receiving packets for 66.8 seconds owing to the HO. As a result, in a highly dynamic handover situation, the throughput is much higher than earlier SCTP designs. Fig.10 compares the throughput of several designs versus the SHD design at high speeds using 10 automobiles as an example.

However, as shown in Fig. 10, mSCTP architecture operates effectively when at least one network is low loaded or has no load at all. For all BS load conditions, the throughput is optimal (4 Mbps), except when the BS is totally loaded (50–100s), in which case the throughput reduces to 2.5 Mbps. The same trend can be seen when looking at the packet loss in Fig. 11. Because the car is still connected to the same BS when the network is crowded and a speed-adaptive strategy is not used, the QoS suffers greatly. When background traffic decreases, the network becomes lightly burdened.

Otherwise, there is no load at all, and the maximum QoS improvement is determined, as shown in the Fig. 10 for the 100–150s interval. Table I concludes all the parameters for the three designs. At the end it is clear that the SHD has the outstanding in both cases with and without background.

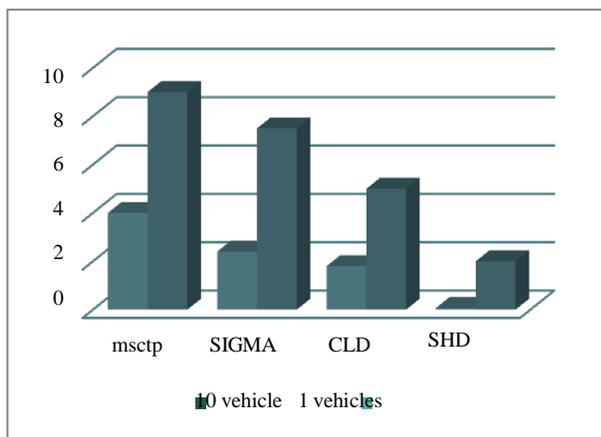


Fig. 10. Throughput of the Background Traffic.

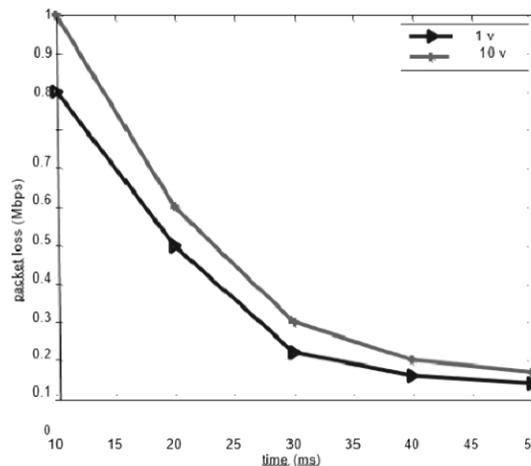


Fig. 11. Packet Losses for Two Cases.

TABLE I. HO DELAY, THROUGHPUT, LOSS, AND AVERAGE SNR, WHEN PERFORMING HANDOVER WITH HIGH SPEED OF 40M/S

Scheme	CL	HO delay	SNR (dB)	Throughput (Mbps)	Loss (Mbps)
SHD	1	0.07	64.3	11.8	0.04
	10	7.5	51.9	9.4	0.07
CLD	1	1.8	45.24	8.43	0.06
	10	5	35.33	7.1	0.1
SIGMA	1	2.4	33.76	5.8	0.12
	10	2	24.5	4.5	0.2

VII. CONCLUSION

Vehicles which normally move among cities at fast speeds have become basic computation in internet communication as IoV, thanks to the rapid growth of communication networks through the Internet. This sort of connection (IoV) may encounter a number of problems that degrade the quality of the Internet connection by lengthening the handover time (HoD). This work presents an approach that uses a cross-layer architecture SHD to dynamically lower the HoD in order to improve connection continuity while dealing with fast-moving data. The suggested architecture has reduced the delay by assisting L4 of the protocol for handover existence, allowing it to complete the handover in advance, resulting in even more secure and lossless vehicle communication. The numbers clearly indicate the improvements in throughput, latency, and packet loss.

ACKNOWLEDGMENT

“This work was funded by the Deanship of Scientific Research at Jouf University under grant No (DSR-2021-02-0360)”.

REFERENCES

- [1] W. Zhidong et al., "Protection settings transmitted over SCTP protocol in intelligent substations," 2018 International Conference on Power System Technology (POWERCON), Guangzhou, 2018, pp. 1439-1444.
- [2] Abdullah, Adel. A & Alzahrani, Ahmad. (2018). A comprehensive survey on handover management for vehicular ad hoc networks based on 5G mobile networks technology. Transactions on Emerging Telecommunications Technologies. 30. 1-19. 10.1002/ett.3546.
- [3] Tuyisenge L., Ayaida M., Tohme S., Afilal LE. (2018) Networks Architectures on the Internet of Vehicles (IoV): Review, Protocols Analysis, Challenges, and Issues. In: Skulimowski A., Sheng Z., Khemiri-Kallel S., Cérin C., Hsu CH. (eds) Internet of Vehicles. Technologies and Services Towards Smart City. IOV 2018. Lecture Notes in Computer Science, vol 11253. Springer, Cham.
- [4] T. T. Dandala, V. Krishnamurthy and R. Alwan, "Internet of Vehicles (IoV) for traffic management," 2017 International Conference on Computer, Communication and Signal Processing (ICCCSP), Chennai, 2017, pp.1-4. doi: 10.1109/ICCCSP.2017.7944096.
- [5] Y. Han, F. Teraoka, "An SCTP Fast Handover Mechanism Using a Single Interface Based on Cross-Layer Architecture". IEICE Transactions, 2009: 2864~2873.
- [6] K. Zhu et al, "Mobility and Handoff Management in Vehicular Networks: A Survey", Wiley InterScience, Communication and Mobile Computing, pp.1-20, 2009.
- [7] S.Fu, M. Atiquzzaman, "Survivability evaluation of SIGMA and mobile IP", Springer, Wireless Communication, 2007.
- [8] Y.S. Chen and C.H. Cheng, "Network Mobility Protocol for Vehicular Ad Hoc Networks", Wireless Communications and Networking Conference, WCNC, IEEE pp.1 – 6, April 2009.
- [9] K. Chiu, R. Hwangy and Y. Chen, "Cross-Layer Design Vehicle- Aided Handover Scheme in VANETs", wireless communications and mobile computing, pp.1–13, 2009.
- [10] S. Fu, M. Atiquzzaman, "Handover latency comparison of SIGMA, FMIPv6, HMIPv6, and FHMIPv6", IEEE GLOBECOM proceeding, Vol. 6, pp. 3809 – 3813, January 2006.
- [11] Shi H., Hamagami T. (2010) Cross-Layer Routing Method for the SCTP with Multihoming MIPv6. In: Hei X.J., Cheung L. (eds) Access Networks. AccessNets 2009. Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering, vol 37. Springer, Berlin, Heidelberg.
- [12] Waqas A. Intiaz, M. Afaq, Mohammad A.U. Babar. (2011) mSCTP Based Decentralized Mobility Framework. (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 2, No.9.
- [13] Alemam, Fatma Alzahra & Nasr, M. & Kishk, Sherif. "Coordinated Handover Signaling and Cross-Layer Adaptation in Heterogeneous Wireless Networking". Mobile Networks and Applications 2019.
- [14] Al Emam, F.A., Nasr, M.E. & Kishk, S.E. "Collaborative cross-layer framework for handover decision in overlay networks". Telecommun Syst 73, 189–203 (2020). <https://doi.org/10.1007/s11235-019-00604-5>.
- [15] Han Y, Teraoka F. SCTPmx: An SCTP Fast Handover Mechanism Using a Single Interface Based on Cross-Layer Architecture. IEICE Transactions, pp. 2864–2873 Vol.E92-B, Sep 2009.
- [16] Liu J, Bi J, Ge Y, Cui X, Ding S and Li Z. A compensation model of cooperative downloading for vehicular network. Trans. Emerging Tel. Tech. 2013; 4:532–543. DOI: 10.1002/ett.2626.
- [17] Jaiganesh B, Ramachandran R. Signaling Cost Analysis of Mobility Management Entities for SIGMA. IJCSNS International Journal of Computer Science and Network Security, VOL.12 No.7, July 2012.
- [18] Reaz A, Roy R, Atiquzzaman M. P-SIGMA: Security Aware Paging In End-To-End Mobility Management Scheme. Journal Wireless Networks (19)8:2049-2065, Nov 2013.
- [19] Fu S, Atiquzzaman M. Architecture and Performance of SIGMA: A Seamless Mobility Architecture for Data Networks. IEEE International Conference on Communications. 2005; 5:3249–3253.
- [20] Fu S, Atiquzzaman M. Handover latency comparison of SIGMA, FMIPv6, HMIPv6, and FHMIPv6. IEEE GLOBECOM Proceeding 2006; 6:3809–3813.
- [21] Chowdhury P, Reaz, Lin TC, Atiquzzaman M. Design Issues for SIGMA: Seamless IP diversity based Generalized Mobility Architecture. Technical Report, Feb. 2006.
- [22] J.Lloret, Z. Ghafoor, B.Rawat, F.Xia. Advances on Network Protocols and Algorithms for Vehicular Ad Hoc Networks. Mobile Netw Appl (18):749–754. Nov 2013. DOI 10.1007/s11036-013-0490-7.
- [23] Budzisz, L., Garcia, J., Brunstrom, A., and Ferr ´ us, R. 2012. A taxonomy and survey of SCTP research. ACM Comput. Surv. 44, 4, Article 18 (August 2012), 36 pages.
- [24] Han Y, Teraoka F. SCTPfx: A Fast Failover Mechanism Based on Cross-layer Architecture in SCTP Multihoming. AINTEC 2008; 113–122, Bangkok, Thailand. DOI 10.1145/2333112.2333113 <http://doi.acm.org/10.1145/2333112.2333113>
- [25] Mugga C, Sun D, Ilie D. Performance Comparison of IPv6 Multihoming and Mobility Protocols. ICN (2014): 166-171.
- [26] Ming C, Shu M, and Hsu TH. PFC: A packet forwarding control scheme for vehicle handover over the ITS networks. Computer Communications 2007; 2815–2826.
- [27] Cen S and Cheng C, Network Mobility Protocol for Vehicular Ad Hoc Networks, Wireless Communications and Networking Conference, IEEE WCNC, pp.1 – 6, April 2009.
- [28] Chiu K, Hwang R, Chen Y. Cross-layer design vehicle assisted handover scheme in VANETs. Wireless Communications and Mobile Computing 2011; 11(7): 916–928.
- [29] Ashutosh D., Henning S., Mobility Protocols and Handover Optimization: Design, Evaluation and Application. ISBN: 978-0-470-74058-3, 476 pages, March 2014, Wiley-IEEE Press.
- [30] Abrougui K, Boukerche A, Ramadan H. Performance evaluation of an efficient fault tolerant service discovery protocol for vehicular networks. Journal of Network and Computer Applications. 35 (2012) 1424–1435.
- [31] Jubara HEI, Ariffin SHS. Evaluation of SIGMA and SCTPmx for High Handover Rate Vehicle. IJACSA 2011; 2(7):169 –173.
- [32] Kumar R, Kaushal S. A Survey of mSCTP for Transport Layer Mobility Management. Journal of Advances in Information Technology, (4)1:20-27, Feb 2013. doi:10.4304/jait.4.1.20-27.
- [33] Pack S. Choi Y. Performance Analysis of Fast Handover in Mobile IPv6 Networks. IFIP International Federation for Information Processing 2003.
- [34] Wen-Kang J., Chia-Yao Chen, and Yaw-Chung Chen, PSFCS: Robust Emergency Communications Supporting High Mobility Based on WiMAX MMR Networks. International Journal of Distributed Sensor Networks, vol. 2014, pp. 1–11, 2014.
- [35] Ciubotaru B, Muntean G, A Quality-Oriented Handover Algorithm for Multimedia Content Delivery to Mobile Users. IGI Global, pp.1-30 (2012). DOI: 10.4018/978-1-61350-107-8.ch001.
- [36] Re ´ n´ e S, Esparza O, Alins J, VSPLIT: A Cross-Layer Architecture for V2I TCP Services Over 802.11. Mobile Netw Appl (18):831–843, 3 Oct 2013. DOI 10.1007/s11036-013-0473-8.
- [37] Chiu K, Hwang R, Chen Y. Cross-layer design vehicle assisted handover scheme in VANETs. Wireless Communications and Mobile Computing 2011; 11(7): 916–928.
- [38] Xiaonan W, Haili H, Hongbin C and Rong Z. A scheme for connecting vehicular networks to the Internet. Trans. Emerging Tel. Tech. Oct 2013. DOI: 10.1002/ett.2743.
- [39] Whaiduzzaman M, Sookhak M, Gani A, Buyya R. A survey on vehicular cloud computing. Journal of Network and Computer Applications, 40 (2014) 325–344.

The Impact of Security and Payment Method On Consumers' Perception of Marketplace in Saudi Arabia

(Case Study on Noon)

Mdawi Alqahtani¹

Department of Business, Umm Al Qura University
P.O. Box 715, Mecca, Saudi Arabia

Marwan Ali Albahar²

Department of Science, Umm Al Qura University
P.O. Box 715, Mecca, Saudi Arabia

Abstract—Digital transformation has been accelerated in recent years, and COVID-19 has resulted in a rise in overall internet spending. Businesses must take measures in order to ensure that customers have a safe and enjoyable online purchasing experience. In this paper, customers' security perceptions regarding the most popular e-commerce applications in Saudi Arabia are explored. Surveys were distributed online via Google Form to 200 participants in total as part of a cross-sectional research design using quantitative methodology. The main findings were related to confirming eight main hypotheses of the research that were related to testing if some factors were important to forming perceived trust by customers. Five factors (trust, security, reputation, benefits, and convenience) were found to have a positive effect, and the remaining three were not (familiarity, size, and usefulness). Finally, this study recommends various actions for practitioners and policymakers to take in order to improve customer perceptions of payment methods and security in Saudi Arabia.

Keywords—Payment security; digital strategy; digital transformation; user security

I. INTRODUCTION

E-commerce is a type of business transaction where goods and services are bought, sold, and given away over the internet. This includes both business-to-business (B2B) and business-to-consumer (B2C) deals. Consumer behavior has changed because of today's fast-paced market and the fierce competition between businesses [1]. Because of this, e-commerce has become a good choice for many businesses. Classified ads, C2C marketplaces, shopping malls, B2C stores, and social media-based online stores are all examples of e-commerce business models in Saudi Arabia, where people buy and sell things online. Because so many people in Saudi Arabia use mobile devices all the time, Noon is one of the best websites for shopping at e-commerce stores in the country [2, 3]. The Noon website sells merchandise. Noon's goal in Saudi Arabia is to provide a new shopping experience as well as simple sales, secure payment, and integrated logistics. As a result, it is critical for the application system to keep data secure and to have a method to ensure that the transaction records it generates are correct because it will be used to address many threats to the payment system [4]. Consumers must have faith in the security of personal data protection for online

remittances for this to happen. Increasingly, people are using smartphones and tablets to do things like make payments and transfer money because they are easy to use and easy to move around. Traditional financial transactions are thought to be more time-consuming and inefficient than mobile financial transactions [5,6].

Customers' perceptions of their personal information and the security of their financial transactions have shifted dramatically in recent years. Mobile applications with a good reputation for security and privacy are used to maintain security and privacy. Customers of mobile applications are concerned about malicious code and unauthorized access infiltrating their applications. They're also worried about others spying on their online activities and stealing their credit card information. Customers' willingness to buy and sell products through mobile commerce applications has been shown to be influenced by privacy and security concerns [6,7,8,9]. Customers are less likely to use an online platform if they are concerned about its security. To better engage with their customers, businesses must first understand how they perceive security and safety when they are online. Because social and cultural factors vary by country, it is critical to investigate customers' perceptions of online platform security and privacy across cultures [10,11]. User behavior shifts in the presence of cybersecurity threats, according to the Protection Motivation Theory (PMT), making it critical for social commerce platforms to protect their customers' personal information and data. As a result, we concentrate our research on customers in Saudi Arabia to investigate if the payment methods used have a positive impact on customer perceptions of security.

Saudi Arabia is widely regarded as a major online shopping market. People buy products using a variety of payment methods and applications all over the world. They are also concerned about the correlation between security, privacy, and payment methods [12, 13]. This motivates us to study the importance of security and privacy for Saudi buyers when it comes to mobile commerce. The objective of this study is to examine Saudi consumers' perspectives on security and privacy concerns when it comes to a popular e-commerce app, namely Noon in Saudi Arabia (KSA). This study used quantitative methodology to determine whether the payment methods that are used have a positive impact on customers' perceptions of

security. We used the most popular e-commerce website in Saudi Arabia, which is Noon. Multiple factors are made up of several sub-factors that helped us analyze the user experience of this e-commerce website. As a result, the main contribution of this research is to better understand customer perceptions of the payment methods and security in Saudi Arabia to improve customer perceptions that can be applied to other countries.

The rest of this paper is organized as follows. Section II presents recent related works on the impact of security and payment methods on consumers' perceptions. We present the problem formulation in Section III. Section IV introduces our proposed methodology in detail. Section V discusses the study setting and participants. Section VI presents the results and findings and some limitations. Finally, we conclude the paper and discuss some directions for future work in Section VII.

II. LITERATURE REVIEW

Numerous studies have shown that security and privacy are critical in mobile commerce applications. Customers are thus more likely to use mobile commerce applications that provide enhanced security and privacy. According to a study conducted in Indonesia by Hidayat et al. [14], trust is a critical factor for Indonesian customers when shopping online. Saprikis and Avlogiaris [15] conducted an empirical study to ascertain the factors that influence consumers' social media shopping behavior. The findings indicate that convenience, reward, and security all play a significant role in consumers' direct purchases via social applications (ICT facilitators of the UTAUT model). Customer satisfaction, according to Taherdoost and Madanian [16], is a critical factor in customer loyalty, which is why they validated an e-service satisfaction model in an e-commerce context. Customer satisfaction is determined to be most strongly influenced by trust, security, performance, and usability. According to Harris et al. [17], a variety of factors contribute to people installing and using mobile applications that violate their privacy and security. In their study, they discovered a correlation between customer trust and perceived security, indicating that customers who perceive more security have both increased trust and a lower perceived risk. According to Ghayoumi [18], m-commerce is also affected by six security factors. Integrity, non-repudiation, authentication, confidentiality, privacy, and availability are some of these factors. According to the study's researchers, security in m-commerce applications is contingent upon these factors. Mahmoud and colleagues [19] attempted to demonstrate the growing popularity of mobile commerce by highlighting the security risks associated with the use of modern devices and high-speed internet. While the study's primary objective was to make recommendations on how to address potential privacy and security concerns raised by the growing use of mobile commerce, it also examined user perceptions of trust in mobile commerce on three major websites: Amazon.com, AliBaba.com, and eBay. The study takes a deductive approach and employs only one method of research. Additionally, the data was analyzed using a 100-respondent random sample. Although an e-commerce environment is more private than a mobile commerce environment, the author asserts that trust in mobile commerce systems remains low due to the privacy and security paradigm. Additionally, this study demonstrates that there is considerable

room for improvement in terms of privacy and integrity, authentication, and security. According to Kumar and colleagues' research [20], while m-commerce applications are gaining popularity in India, users remain wary of them for a variety of reasons. Security and payment issues with these mobile commerce applications have been cited as significant factors. Consumers believe that mobile commerce applications are constantly vulnerable to hacking and phishing attacks and that they cannot trust any of them with their credit cards or even personal information. They are wary of using these apps because they lack trust in third-party websites to process their payments. Venkatesh et al. [21] also investigated customers' privacy concerns when making online purchases. It was discovered that among the recommendations included in the study were retailers' and other customers' preferences for products that were related to one another. According to a survey, online purchases are moderately influenced by recommendations and product-relatedness. Closely associated products with privacy enablers did not affect online purchase intentions. As Gurung et al. [22] discovered when they investigated online shoppers' security and privacy concerns. These concerns, they believe, influence customers' perceptions of risk. The relationship between privacy and security was also examined using organized conduct. The study's findings indicate that privacy and security concerns may influence risk assessment and awareness. Privacy and security concerns rank second, with trust ranking first. Additionally, individuals' mental states are affected by their perceptions of risk and trust. Ali et al. [23] also examined privacy to determine whether it could be used to deduce users' attitudes toward mobile app security. Värzaru et al. [24] used a reworked version of the technology acceptance model to deduce the factors influencing post-COVID behavioral intention and consumer satisfaction. According to the researchers, consumer intention was positively influenced by perceived usefulness and ease of use. D'Adamo and colleagues [7] discovered that in the post-pandemic era, European consumers are concerned about online security. According to the findings of this study, consumers in Europe have varying levels of concern about e-commerce security. According to Hussien et al. [25], customers and electronic marketplaces can benefit from agent software that provides client-side security to improve marketplace performance. For example, Chen et al. [26] proposed a forensic model that may aid in detecting abnormal system behavior. Numerous studies on e-commerce in Saudi Arabia have been conducted. Between 2013 and 2016, a long-term study by Miao & Tran [27] discovered a significant difference between SMEs' initial adoption of e-commerce and their intention to institutionalize it. Nachar [28] asserts that an e-commerce platform's ease of use and usefulness are statistically significant predictors of customers' willingness to shop online. According to Al-Empirical in Ayed's study [29], customer care, product selection, convenience, personality, and website customization all contribute to e-commerce customer loyalty. Saeed [30] conducted an empirical study of Saudi Arabian expats' adoption of e-commerce and discovered to enhance user interface usability, it is necessary to take cultural differences into account during the technical design stage. Alotaibi [31] found no significant differences in m-commerce customer loyalty by gender, age, or prior experience.

According to Razi et al. [32], participation in social commerce by students has a beneficial effect on purchase intentions and behavior.

According to a review of the literature, geographical and cultural factors influence users' attitudes toward online shopping. However, no comprehensive study of online customers has been conducted in Saudi Arabia. Due to the importance of user motivation and perception in technological adoption, there is a knowledge gap regarding how Saudi consumers perceive the security aspects of e-commerce applications.

III. PROBLEM FORMULATION

Certain studies focus exclusively on the relationship between factors influencing marketplace trust, and only infrequently do they examine the relationship between factors influencing payment system trust in developing countries. In the end, a study by Kim et al. [33] looks at whether trust is affected by factors like reputation, privacy, size, security, benefits, and convenience. It also looks at how this trust affects purchases made with EPS, credit cards, or cash on delivery (COD). It is regarded as user-friendly and useful. Security, usability, trustworthiness, interoperability, common issues, and extra services all affect how well electronic payment systems work. These six factors all affect how well electronic payment systems work. Mutual trust between merchants and customers is required when conducting online shopping, based on the six factors listed above. In [34], the author asserts that the value of security and trust cannot be overstated. Privacy refers to the right to keep one's personal information private. Additionally, privacy is defined as the capacity to manage personal information that is required and used by third parties [35]. If you want to buy something on the internet, you must be willing to give out your personal information before you do so [35]. When it comes to interpersonal relationships, humans prefer to maintain their privacy.

In e-commerce, privacy refers to how willing people are to give out personal information over the internet before they buy something [35]. The term "internet privacy" includes a lot of different things, like data, choices, and sharing with e-commerce service providers. As Belanger stated [36], consumers also want to know that the information they give is safe and lawful. The right to privacy of individuals is extremely well protected. When customers shop online, they provide sellers with extremely detailed information. Consumers who place a high value on privacy often give internet service providers inaccurate or incomplete information. It is possible to take advantage of the privacy settings on a website. In other words, the more confident a user is in a website's ability to protect their information, the greater their trust on the website.

A. Hypothesis

Trust is significantly influenced by people's perceptions of their reputation, privacy, size, security, benefits, website usability, and convenience.

- According to [33], a payment system's perceived user friendliness and usefulness is critical. Security, usability, trustworthiness, interoperability, common

issues, and additional services are all factors that have an impact on the performance of electronic payment systems. It's important to think about safety first. As shown by the preceding six factors, online shoppers and merchants need to put in a lot of work to build trust in one another.

1) *Perceived privacy has a positive effect on trust:*

According to the authors in [8], privacy is a method of protecting one's identity. The ability to retain one's personal information is defined as "privacy" in this definition. Also included in the definition of privacy is an individual's ability to control the extent to which their personal information is required and used by third parties [35]. Privacy online can be defined as consumers' willingness to share personal information before making a purchase [35]. Humans, like other people, have a standard for how much privacy they want. In e-commerce, privacy is defined as consumers' willingness to provide information via the internet before making a purchase [35]. Concerns about privacy on the internet include "spam," "data," "choices," and e-commerce service providers sharing information. Customers also want assurances that the information they provide will be restricted and regulated by the person concerned [36]. Everyone has a right to have their personal information kept private. Customers in the e-commerce industry are extremely picky about the information they divulge to merchants. Internet service providers are more likely to receive incomplete information from consumers who care about their privacy. When you give your personal information to a website, you run the risk of it being misused. The more trust is placed in an address's ability to protect personal information, the more confident that person is in that address's ability to protect that information.

2) *In general, security increases when people feel safe:*

Security is a significant control issue for businesses that conduct e-commerce. When consumers are involved in electronic transmission, data relating to e-commerce, such as buyer and seller data, must be kept confidential. Additionally, the transmitted data must be safeguarded against modification or alteration by anyone other than the sender [36]. According to [37], security can be defined broadly as the absence of danger. This understanding is comprehensive and encompasses an individual's sense of protection against both intentional and unintentional crimes, such as natural disasters. A security threat is defined as a situation, condition, or event that poses a risk of causing damage to data or networks, which can take the form of data destruction, leakage, alteration, or misuse. Consumer security concerns can be addressed in e-commerce using protection technology. When these technologies are used, they are classified as security features. According to [9], security can be classified into four categories based on security holes: physical security (physical security), personnel security (personnel security), data security (data security), and media and communication techniques (communications). Security in operations refers to the policies

and procedures that govern the establishment and management of security systems, as well as post-attack recovery procedures. The management of the online payment system's security can be viewed through the lens of risk management. Authors in [37] recommended employing the "Risk Management Model" when confronted with threats (managing threats). Risk consists of three components: assets, vulnerabilities, and threats. E-commerce network security that incorporates features such as guarantees, contracts, or other procedures ensures the existence and proper operation of payment security. Someone who has a high perception of structural assurance will fervently believe that internet technology (e.x. data encryption) protects in such a way that online transactions are safe. Consumers are protected from financial and personal loss through encryption, legal protection, and technological safeguards. In addition, the authors in [37] stated that security guarantees could be integrated into e-commerce sites through collaboration with third parties with a strong reputation in network security and who provide internet security assurance standards via web assurance seals. Consumers who feel secure in the online environment are more likely to trust websites that offer electronic commerce services than those who believe the internet is unsafe because they do not believe e-commerce sites offer adequate protection.

3) *Perceived benefits are considered when trust is enhanced:* According to [36,37], usefulness is the likelihood that a specific application will be used by potential users to make their work tasks easier. Results will be obtained more quickly and satisfactorily as a result of the product's simplified performance when used in conjunction with the new technology. Internet banking services can boost productivity by increasing people's perceptions of the benefits these services provide. Increased productivity, improved performance, and improved process efficiency can all be used to determine what people think about the benefits of technology.

4) *Consumer trust in a website is increased as a result of familiarity with it:* MAQABLEH et al. [39] study and observe consumer behavior and perceptions of security and trust in e-payment systems based on the proximity of the customer to the website. In addition, the authors in [39] identified a slew of determining factors. There was also a tendency to trust, as well as internet experience, personal innovation, and habit. Third-party involvement, payment system intention, enjoyment, risk aversion, and trustworthiness can all be found by looking at the variables that connect them. According to investigation made in [33], after a customer's first visit to a company's website, the level of consumer confidence in that company's website was measured. According to the findings of the investigation, consumers' perceptions of the company's reputation and willingness to improve products and services have a direct impact on consumer confidence. In addition to the other factors, it is thought that controls for usability, ease of use, and security have a big impact on trust.

5) *Perceived convenience has a positive effect on trust:* According to [37,38], ease of use can be defined as the extent to which a person believes that using technology will be free of effort on his or her part. As implied by the definition, ease of perception is a belief about the decision-making process that is experienced by the individual. Using an information system is more likely to occur if a consumer believes it is simple to use and understand. As identified by [36-37-38], the dimensions of perceived ease are as follows: ease of learning (easy to learn), ease of use (easy to use), clear and understandable (straightforward and easy to understand), and the ability to become skillful (becoming skilled).

6) *Perceived trust has a positive impact on purchase intention when EPS is employed:* Consumers' online behavior is heavily influenced by their level of trust in the companies they do business with, which is why trust is such an important consideration in electronic commerce. One's social standing rises as a result, and one can spend money in the market. According to [8-34-35-36], trust is based on the trustworthiness of the parties involved in the transaction, specifically electronic payments and cross-border trade. Trust in other parties and the use of regulatory control mechanisms were found to be the most important factors in determining the level of trust in transactions. Both variables have objective and subjective components. A lack of trust is a direct indicator of attitude and behavior because of the dynamic nature of cyberspace's high uncertainty and constant change. Trust can also be defined as a person's belief in the ability of others to be trusted, which is based on perceived integrity, benevolence, and competence. The most basic definition of trust is the belief that others will not take advantage of you and that the vendor will deliver on what they have promised. Online shopping relies heavily on trust, which is a significant factor in e-commerce. For e-commerce to work properly, trust and security are two of the most critical constructs, customers tend to have a higher level of trust in e-commerce websites with higher quality content. In a developing country, establishing trust in a new environment is a challenging task that is essential to influencing consumer attitudes [4,5,30,31,33].

7) *Perceived trust influences cash on delivery (or/and) credit card purchases:* Tsiakis and Sthephanides [40] argued that trust and security are two of the most important factors to consider when developing an electronic payment system. According to Kim et al. [33], user convenience and usefulness are important factors to consider when choosing a payment system for their needs. In fact, the ability to feel safe and confident in a company's products and services is critical to attracting and retaining customers.

IV. PROPOSED METHOD

This study's quantitative design collects data from a pool of participants one at a time. Customers of the Saudi Arabian online shopping platform Noon were chosen as the population and sample for this study based on the researchers' judgment of what they should buy. Hair et al. [41] stated that the number of samples in PLS-SEM research must be five times the number

of questions in the questionnaire. As a result, this study's questionnaires contain 5 x 40 questions, yielding a total of 200 respondents. The research questionnaire had closed questions with one of five measurement scales for each variable (Likert). This is done using Google Form and explains what will be done to respondents via social media such as WhatsApp, Instagram, and other social media.

The components of this study were adapted from Maqableh et al. [39] findings on trust behavior and online shopping payment methods for shoppers in Saudi Arabia. The Likert scale was used to measure all constructs in this study, with 1 representing "strongly disagree" and 5 representing "strongly agree." Table I summarizes the results of the fittest for the overall PLS-SEM model. With the Good of Model Fit (GOM) metric, the structural model testing phase can be done as follows:

TABLE I. MODEL FIT GOODNESS

Goodness Model of Fi	Original Value (Saturated Model)	Estimated Model	Note
d_U LS	4.23	9.378	Model Fit
SRMR	0.05	0.088	Model Fit
d_G	1.63	1.80	Model Fit

The Standardized Root Mean Square Residual (SRMR) graph illustrates the amount of error associated with predicting the independent variable's effect on the dependent variable in question. According to the definitions of d_U LS and d_G, a representative research model must have a value greater than 0.05 (if the 95 percent confidence interval is used) or greater than 0.01 for the study's smaller initial estimate (if using a 99

percent confidence interval). In other words, the research model's residual distribution is quite small. Validity is established when the square root of the average variance (AVE) value has a loading factor greater than 0.5, and reliability is established when the composite reliability value is greater than 0.7 or when Conbach's Alpha has a loading factor of 0.6.

V. STUDY SETTING AND PARTICIPANTS

Data points from the distribution of questionnaires were collected using Google's non-probability form method, and these data points can be used to generate additional research data. The following are the characteristics of the 200 participants in the survey (see Table II). Noon customers are predominantly female, with 80% of respondents to this study's questionnaire distribution reporting a shopping frequency of more than 19 times per month, according to the results of the study. Customers who are the most active on Noon fall into this category. Using a correlation coefficient, it was discovered that the effects of total reputation perception, privacy perception, scale perception, safeguard perception, perceived usefulness, user-friendly perception, and trust perception were all increased by 74.1%. In EPS, the variable assessing trust perception accounted for 42.4% of the total. Fixed trust perception of factual purchases made with credit cards accounted for 15% of the total, while variable trust perception of factual purchases made with cash-on-delivery accounted for another 24% of the total (see Table III). Finally, after figuring out the coefficients of determination for each parameter, we conduct the experiment to determine the validity of the hypothesis (Fig. 1).

TABLE II. DEMOGRAPHIC INFORMATION OF RESEARCH RESPONDENTS

Type of Characteristic	Characteristic	Total	Percentage
Sex	Male	40	20%
	Female	160	80%
Age	14 – 20 years old	102	51%
	21 – 30 years old	60	30%
	40 – 50 years old	18	9%
	> 50 years old	20	10%
Occupation	Students	100	59,27%
	Public Sector Employees	43	17,34%
	Private Sector Employees	40	18,95%
	Enterprise Employees	17	4,44%
Income	< RS 6,000	124	69,35%
	RS 7,000 – 12,000	50	20,16%
	RS 13,000 – 17,000	15	6,05%
	> RS 20,000	11	4,44%
Shopping Frequency	< 4 times	70	34,68%
	8 times	50	26,21%
	12 times	20	8,06%
	>19 times	60	31,05%
Shopping Cost	< RS 4,000	140	75,40%
	RS 5,000	48	19,76%
	RS 7,000	4	1,61%
	RS 15,000	8	3,23%

TABLE III. COEFFICIENT OF DETERMINATION

Model	R Square
Trust is significantly influenced by people's perceptions of their reputation, privacy	0.74
Purchase with the intention of utilizing EPS	0.428
Purchase made with credit card	0.015
Actual purchase made with on-delivery payment	0.024

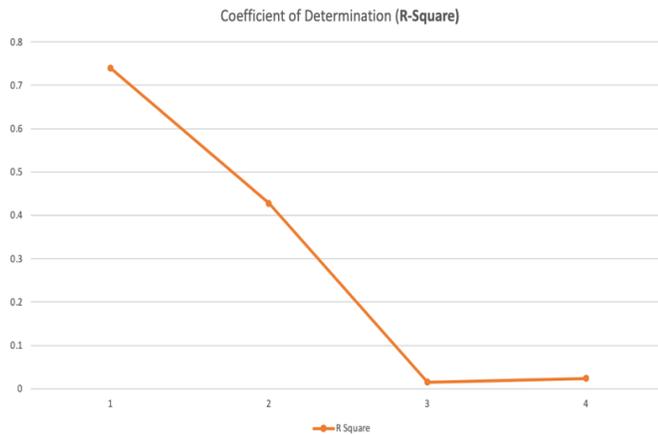


Fig. 1. Coefficient of Determination (R-square).

VI. RESULT AND FINDINGS

These findings demonstrate that belief-formers have an impact on online shopping on the Noon application in Saudi Arabia, which is decided by factors such as safety, benefits, and convenience, all of which have a statistically significant impact. According to the authors in [12,33], high-quality e-commerce sites have a greater perception of trust from their customers, and the exceptional measures taken to earn the trust of customers will shape consumer attitudes in a developing country.

In this study, the researchers discovered that clients who shop online through the Noon application are more likely to purchase things from Noon when they pay with an electronic

funds transfer (EPS). According to the findings of Tsiakis and Sthephanides [40], credence and safeguard are the most important and vital components for electronic payment systems that are used as a tool in financial operations. Furthermore, according to the authors in [34], security and trust play crucial roles in recruiting and retaining customers. Also demonstrated in this study is that the perceived safety, benefits, and convenience of shopping online at Noon in Saudi Arabia have an impact on trust perception, and that using EPS is the most influential factor in trust perception when it comes to purchasing online. As a result, marketplace service providers such as Noon may be able to develop confidence by emphasizing the safety, benefits, and convenience of online shopping. In the long run, service providers may be able to design payment mechanisms that are compatible with EPS. The likelihood of consumers making purchases online at Noon will increase as their degree of confidence increases, as will the likelihood of consumers using EPS payment options (Fig. 2). The online services provided by Noon in Saudi Arabia will be directly recommended to clients who have expressed satisfaction with the company's trustworthiness, security, and convenience of payment (see Table IV).

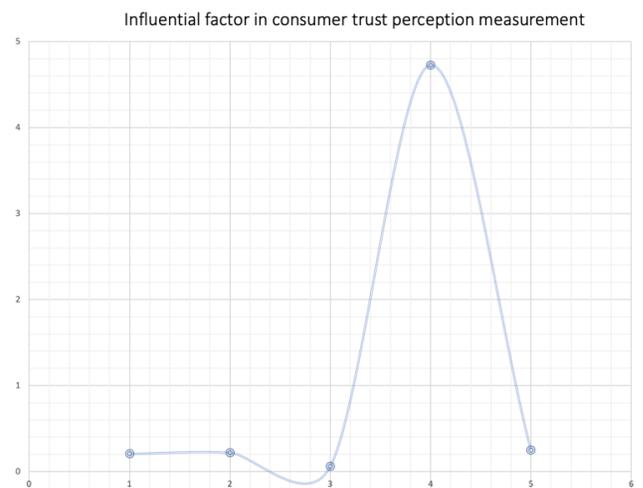


Fig. 2. Influential Factor in Consumer Trust Perception Measurement.

TABLE IV. HYPOTHESIS TEST

Model	Original Sample	Mean of the ample	Standard Deviation	T-Value	P-Value	Note
Reputation, privacy, size, security, benefits, website usability, and convenience all influence trust.	0.155	0.151	0.070	2.224	0.021	significant
A favorable perception of the firm's size increases trust.	0.013	0.002	0.057	0.047	0.951	Not significant
People trust more when they feel safe.	0.411	0.452	0.057	7.749	0.000	significant
Perceived benefits increase trust.	0.112	0.165	0.055	2.560	0.015	significant
Familiarity increases consumer trust in a website.	0.030	0.020	0.083	0.363	0.711	Not significant
Convenience has a positive effect on trust.	0.156	0.182	0.081	2.227	0.039	significant
When EPS is used, perceived trust positively impacts purchase intention.	0.662	0.665	0.031	20.645	0.000	significant
Trust perception affects cash on delivery / credit card purchases	0.127	0.131	0.062	1.986	0.042	significant

VII. CONCLUSION

In recent years, digital transformation has accelerated, and COVID-19 has resulted in an increase in overall internet spending. Businesses must take precautions to ensure that their customers have a safe and enjoyable online shopping experience. This paper investigates customers' security perceptions of the most popular e-commerce applications in Saudi Arabia. As part of a cross-sectional research design employing quantitative methodology, surveys were distributed online via Google Form to a total of 200 participants. The main findings were related to confirming the research's eight main hypotheses, which were related to testing whether some factors were important in forming perceived trust by customers. Five factors were discovered to have a positive effect (trust, security, reputation, benefits, and convenience), while the remaining three did not (familiarity, size, and usefulness). Based on our findings, trust is a multidimensional construct comprised of reputation, privacy, size, benefits, security, benefits, familiarity with the web, and ease. As demonstrated by this result, the p-value does not apply to all indicators with a loading of less than 0.04 on the latent variable. This study suggests several actions that practitioners and policymakers can take to improve customer perceptions of payment methods and security in Saudi Arabia.

REFERENCES

- [1] Reychav, I.; Beeri, R.; Balapour, A.; Raban, D.R.; Sabherwal, R.; Azuri, J. How reliable are self-assessments using mobile technology in healthcare? The effects of technology identity and self-efficacy. *Comput. Hum. Behav.* 2019, 91, 52–61.
- [2] Taneja, B. The Digital Edge for M-Commerce to Replace E-Commerce. In *Emerging Challenges, Solutions, and Best Practices for Digital Enterprise Transformation*; IGI Global: Hershey, PA, USA, 2021; pp. 299–318.
- [3] Ortiz, J. The global environment through the SLEPT framework. *Int. J. Bus. Glob.* 2010, 5, 475–492.
- [4] Chaffey, D.; Edmundson-Bird, D.; Hemphill, T. *Digital Business and E-Commerce Management*; Pearson: London, UK, 2019.
- [5] Saeed, S.; Bolívar, M.P.R.; Thurasamy, R. *Pandemic, Lockdown, and Digital Transformation*; Springer: Berlin/Heidelberg, Germany, 2021.
- [6] Niranjanamurthy, M.; Kavyashree, N.; Chahar, S.J.D. Analysis of E-Commerce and M-Commerce: Advantages, Limitations and Security issues. *Int. J. Adv. Res. Comput. Commun. Eng.* 2013, 2, 2360–2370.
- [7] D'Adamo, I.; González-Sánchez, R.; Medina-Salgado, M.S.; Settembre-Blundo, D. E-Commerce Calls for Cyber-Security and Sustainability: How European Citizens Look for a Trusted Online Environment. *Sustainability* 2021, 13, 6752.
- [8] Pabian, A.; Pabian, B.; Reformat, B. E-Customer Security as a Social Value in the Sphere of Sustainability. *Sustainability* 2020, 12, 10590.
- [9] Fathima, K.; Balaji, D.S. Enhancing Security in M-Commerce Transactions Enhancing Security in M-Commerce Transactions. *Ann. Rom. Soc. Cell Biol.* 2021, 25, 3915–3921.
- [10] Clemons, E.K.; Wilson, J.; Matt, C.; Hess, T.; Ren, F.; Jin, F.; Koh, N.S. Global Differences in Online Shopping Behavior: Understanding Factors Leading to Trust. *J. Manag. Inf. Syst.* 2016, 33, 1117–1148.
- [11] Mohammed, Z.A.; Tejay, G.P. Examining privacy concerns and ecommerce adoption in developing countries: The impact of culture in shaping individuals' perceptions toward technology. *Comput. Secur.* 2017, 67, 254–265.
- [12] Lee, D.; LaRose, R.; Rifon, N. Keeping our network safe: A model of online protection behavior. *Behav. Inf. Technol.* 2008, 27, 445–454.
- [13] Alotaibi, A.R.; Faleel, J. Investigating the preferred methods of payment for online shopping by Saudi Customers. *PalArch's J. Archaeol. Egypt Egyptol.* 2021, 18, 1041–1051.
- [14] Hidayat, A.; Wijaya, T.; Ishak, A.; Catyanadika, P.E. Consumer Trust as the Antecedent of Online Consumer Purchase Decision Information 2021, 12, 145.
- [15] Saprikis, V.; Avlogiaris, G. Factors That Determine the Adoption Intention of Direct Mobile Purchases through Social Media Apps. *Information* 2021, 12, 449.
- [16] Taherdoost, H.; Madanchian, M. Empirical Modeling of Customer Satisfaction for E-Services in Cross-Border E-Commerce Electronics 2021, 10, 1547.
- [17] Harris, M.A.; Brookshire, R.; Chin, A.G. Identifying factors influencing consumers' intent to install mobile applications. *Int. J. Inf. Manag.* 2016, 36, 441–450.
- [18] Ghayoumi, M. Review of Security and Privacy Issues in e-Commerce. In *Proceedings of the International Conference on Learning, e-Business, Enterprise Information Systems, and e-Government*, Las Vegas, NV, USA, 25–28 July 2016; p. 156.
- [19] Mahmoud, M.A.; Khrais, L.; AlOlayan, R.M.; Alkaabi, A.M.; Suwaidi, S.Q.A.; Alghamdi, B.A.; Aljuwaie, H.F. Consumers Trust, Privacy and Security Issues on Mobile Commerce Websites. *Mod. Appl. Sci.* 2019, 13, p21.
- [20] Kumar, U.; Gope, A.K.; Singh, S. Emerging Challenges and Opportunities of Mobile Commerce in India: A Study on Societal Perspective. *Comput. Trends J. Emerg. Trends Inf. Technol.* 2016, 6.
- [21] Venkatesh, V.; Hoehle, H.; Aloysius, J.A.; Nikkiah, H.R. Being at the cutting edge of online shopping: Role of recommendations and discounts on privacy perceptions. *Comput. Hum. Behav.* 2021, 121, 106785.
- [22] Gurung, A.; Raja, M.K. Online privacy and security concerns of consumers. *Inf. Comput. Secur.* 2016, 24, 348–371.
- [23] Ali, B.J. Impact of COVID-19 on consumer buying behavior toward online shopping in Iraq. *Econ. Stud. J.* 2020, 18, 267–280.
- [24] Värzaru, A.A.; Bocean, C.G.; Rotea, C.C.; Budică-Iacob, A.-F. Assessing Antecedents of Behavioral Intention to Use Mobile Technologies in E-Commerce. *Electronics* 2021, 10, 2231.
- [25] Hussien, F.T.A.; Rahma, A.M.S.; Wahab, H.B.A. Design and implement a new secure prototype structure of e-commerce system. *Int. J. Electr. Comput. Eng.* 2022, 12, 560–571.
- [26] Chen, C.-M.; Cai, Z.-X.; Wen, D.-W. Designing and Evaluating an Automatic Forensic Model for Fast Response of Cross-Border E-Commerce Security Incidents. *J. Glob. Inf. Manag.* 2022, 30, 1–19. [CrossRef].
- [27] Miao, J.J.; Tran, Q.D. Study on e-commerce adoption in SMEs under the institutional perspective: The case of Saudi Arabia. *Int. J. E-Adopt. (IJE)* 2018, 10, 53–72.
- [28] Nachar, M. Factors that Predict the Adoption of Online Shopping in Saudi Arabia. Ph.D. Thesis, Walden University, Columbia, MD, USA, 16 April 2019.
- [29] Al-Ayed, S. The impact of e-commerce drivers on e-customer loyalty: Evidence from KSA. *Int. J. Data Netw. Sci.* 2022, 6, 73–80.
- [30] Saeed, S. Digital Business adoption and customer segmentation: An exploratory study of the expatriate community in Saudi Arabia. *ICIC Express Letter.* 2019, 13, 133–139.
- [31] Alotaibi, R.S. Understanding customer loyalty of M-commerce applications in Saudi Arabia. *Int. Trans. J. Eng. Manag. Appl. Sci. Technol.* 2021, 12, 1–12.
- [32] Razi, M.J.M.; Sarabdeen, M.; Tamrin, M.I.M.; Kijas, A.C.M. Influencing Factors of Social Commerce Behavior in Saudi Arabia. In *Proceedings of the 2019 International Conference on Computer and Information Sciences (ICCIS)*, Sakaka, Saudi Arabia, 3–4 April 2019; pp. 1–4.
- [33] Kim, C.; Tao, W.; Shin, N. and Kim, K.S. An Empirical Study of Customers' Perceptions of Security and Trust in E-Payment Systems. *Electronic Commerce Research and Applications*, 9, 84-95. use mobile payment. *Computers in Human Behavior*, 26, pp.310–32, 2010.
- [34] Nuri, H. A Study of Role of the Factors Influencing the Acceptance of E-Banking. No. 3, 521-525, 2014.

- [35] Ackerman, S. and Davis, D. T. Jr. Privacy and security issues in e-commerce. In the New Economy Handbook, pages. 911–930. Academic Press/ Elsevier, 2003.
- [36] Belanger, F., Hiller, J. S., & Smith, W. J. Trustworthiness in electronic commerce: the role of privacy, security, and site attributes. *The Journal of Strategic Information Systems*, 11(3-4), 245-270,2002.
- [37] Liu, L., & Shi, W. Trust and reputation management. *IEEE Internet Computing*, 14(5), 10-13, 2010.
- [38] Montibeller, G., Franco, L. A., & Carreras, A. A Risk Analysis Framework for Prioritizing and Managing Biosecurity Threats. In Risk Analysis (Vol. 40, Issue 11, pp. 2462–2477, 2020).
- [39] MAQABLEH, M. *Analysis and design security primitives based on chaotic systems for ecommerce*, 2012. (Doctoral dissertation, Durham University).
- [40] Tsiakis, T., & Sthephanides, G. The concept of security and trust in electronic payments. *Computers & Security*, 24(1), 10-15, 2005.
- [41] Hair, F. Jr, Sarstedt, J., Hopkins, M. , L. and Kuppelwieser, G. , V., "Partial least squares structural equation modeling (PLS-SEM): An emerging tool in business research", *European Business Review*, Vol. 26 No. 2, pp. 106-121, 2014. <https://doi.org/10.1108/EBR-10-2013-0128>, 1989.

Individual Risk Classification of Crime Groups using Ensemble Classifier Method

Ardhito P. Anggana, Amalia Zahra

Computer Science Department, BINUS Graduate Program
Master of Computer Science, Bina Nusantara University, Jakarta, Indonesia 11480

Abstract—The most significant challenge for humanity worldwide to crime, especially terrorist attacks, should be considered. Determining the priority scale for anticipating individual terrorist groups is not easy and will significantly affect work activities and subsequent decision-making measures. Priority scale determination decisions should be made carefully so team members cannot choose the desired priority target. Determining the exact priority scale for a target can be influenced by several factors, such as desire factors and ability factors, using Dataset Intelligence. This research aims to find out the ability of each target and pattern to be carried out. Based on this problem, the study used the K-Nearest Neighbor (KNN), Naïve Bayes (NB), Decision Tree (DT), and Ensemble Bagging methods. Each of these algorithms has its characteristics; This classification technique can group priority targets according to their similarities, abilities, and desires. The value of each method used can be used as a reference to determine the correct group information for officers to determine the next steps. The study obtained a maximum accuracy of 70.25% using the Ensemble Bagging-Backward Elimination-K-Nearest Neighbor (KNN) classification method using 20 features. The results showed tests conducted and final analysis and conclusions based on accuracy and recall performance. The precision performance revealed that the Ensemble Bagged KNN, more precisely than KNN, Naïve Bayes, Decision Tree, and Bagging Naïve Bayes and Bagging Decision Tree. The KNN Bagging ensemble model can add accuracy, map individuals, and detect who should be intensely monitored based on predictive results.

Keywords—Priority scale; data mining; classification; ensemble classifier method

I. INTRODUCTION

The crime of terror is one of the most challenging threats to the global community. Its heterogeneous and complex nature has fostered an increasing interest in the scientific community, primarily to inform policy-oriented measures. The recent availability of large data sets, the diffusion of powerful machines, and advances in mathematical modeling techniques have contributed to the development of several approaches to the study of terrorism [1]. However, it is alarming that the sheer abundance of data makes it nearly impossible for authorities to examine every person, conversation thread, or social media post to classify whether they are linked to terrorism or contain elements of terrorist activity [2].

Forms of criminal acts spread in the corridors of terrorism can be intimidation and threats, murder, persecution, bombings, detonation, arson, kidnapping, hostage-taking, and piracy. The impact of these forms of terror is very diverse,

including the onset of panic, feelings of fear/intimidation, worry, loss of property, incision, and even death. Current hidden or missing link prediction models based on network analysis models rely on machine learning techniques to improve model performance in terms of prediction accuracy and computing power [3]. In addition, with the increasing use of computerized systems to track crimes, computer data analysts have begun helping law enforcement officers and analysts to speed up the crime resolution process [4].

Determining parameters is one of the most vital things in the division based on desire and ability factors to overcome how the data mining process can classify the individual risk of criminal groups to determine priority scales and predict the priority scale in each individual as well as how the comparison of trial results from the model used can predict the priority scale. Determining these parameters is not easy for indicated people and will affect the person's activities indicated and included in the subsequent investigation. Trend analysis is a challenging task because crime data relies heavily on timing. Any data collected around criminal behavior and crime types can repeatedly change during the investigation [5]. In this study, the priority scale in this grouping was divided into three priority scales. With the use of priority scales, we can find out the abilities and desires of the person, indicating whether to commit a criminal act or not.

This research will develop previous research [6]. It presents the best machine learning models that can be used on terrorism-related data to predict terrorist groups most accurately. The decision to take a subset of data analysis was to help overcome the limitations. However, this is still possible to collect, record, and process data using the Decision Tree, Naïve Bayes, K-Nearest Neighbour, and Ensemble Bagging approaches and use optimal classification in analyzing individual data. Datasets using intelligence data with data used amounted to 1088 data with 21 attributes. Of the 1088 data, are people indicated to be committing a crime of terror or related to the act. However, in this study, researchers wanted to determine which attributes affect and do not affect the process with the four algorithms mentioned earlier in the feature selection process. So that later it can be used to form a reliable model in knowing the patterns of individuals who have the possibility of entering the Green, yellow, and orange priority scale using the Ensemble Classifier Method Risk Classification Model [7] because it is theoretically and empirically proven to provide much better performance than Single Learner [8], [9]. This research is limited to RapidMiner as a tool, data using intelligence data, not doing the data imbalance data process.

II. THEORY

In this study, individual risk in the context of terrorism provides a unique opportunity to holistically consider risk factors rather than the individual critical factors often given in analysis [10]. Furthermore, understanding the modus operandi of each terrorist group provides a vast advantage to counter-terrorist institutions so that the necessary steps can be taken first to address the threat posed by those groups [6]. The machine learning approach can solve problems by finding a suitable algorithmic model and is better at generating predictive values from an input variable [7], and has four categories that are generally applied to the concept of data mining [11] is supervised learning, unsupervised learning, semi-supervised learning, and active learning.

In this study, the main categories of Machine Learning used utilizing existing data to perform classification [12]. Such models to perform introduction/classification/prediction is used in crime analysis [13]. When solving problems, no algorithm that provides the desired quality is proposed to use a composition or ensemble algorithm [14]. Data input for classification is a collection of records [15], where x is a set of attributes and y is a particular attribute. Classification models are helpful for descriptive encodings for distinguishing objects from different classes [16] and predictive modeling to predict class labels from unknown records [17]. Classification algorithms will produce patterns or rules that can be used to predict classes. Some of them are Naïve Bayes Classifier[18], Support Vector Machine (SVM), Decision Tree, K-Nearest Neighbor (KNN), and Random Forest Classifier.

The decision tree [14] is one of the exciting classification algorithms for taking measurements using a tree structure consisting of a collection of decision nodes connected by branches from the decision root to the leaf node to produce new decisions until they finally find the correct decision (leaf node) [19]. Data testing is conducted at each decision node[9] to separate datasets into subsets based on data homogeneity. Generally, the Decision Tree method used in modeling is Decision Tree CART. The CART algorithm [7] can build classification and regression modeling using the Gini Index [20] for the attribute selection process. Criteria determine the model of the decision tree formed, which is measured using the formula Entropy.

$$H = \sum P_k \log_2(P_k) \quad (1)$$

K-Nearest Neighbor (KNN), One classification method that can train the model without using parameters (non-parametric) [21] by classifying the object with the most vote values of each predefined object. The technique that can be used to measure the distance between two points or tuples of them is the Euclidean distance technique. Let us say point X is $X_1 = (x_1, x_2, \dots, x_n)$, Then point Y is $Y_1 = (y_1, y_2, \dots, y_n)$. Then the measurement formula used.

$$\text{dist} (d) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \dots + (x_n - y_n)^2} \quad (2)$$

The Naive Bayes algorithm is a supervised learning approach used for classifying to predicting target variables. Generally, classification techniques predict labeled classes based on attributes by looking for significant correlations

between input and output variables. Naïve Bayes can be a simple probabilistic classifier that can build modeling on large datasets without estimating complex parameters [22]. The basic formula used in Bayes' theory.

$$P = \left(Y | X = \frac{P(Y) \times P(Y | X)}{P(X)} \right) \quad (3)$$

Ensemble Classifier Method is a diverse concept of modeling methods used to solve the problem of base learners by developing and combining a set of hypotheses to correct training data weaknesses using a single-learners approach [7]. This method is generally used in classification to build a model, including bagging, boosting, and stacking by reducing errors and optimizing accuracy results to be better than the base classifier itself [7]. The basic formula used in Ensemble Classifier's theory is.

$$F(x) = c_0 + \sum_{m=1}^M c_m T_m(x) \quad (4)$$

In this study, the algorithm used as a meta-classifier in the Ensemble Classification method is a Bagging algorithm. This simple ensemble meta-algorithm learning method helps reduce variance and improve the prediction and stability of feature selection [23]. By attaching each Ensemble basic learners to a subset of instances, by size n , tasted with repetition of instances n available. As a result, base learners will have a low statistical correlation, improving the Ensemble's predictive performance [24] and discarding error change segments [25] where individual algorithm errors are compensated reciprocally [26]. Furthermore, it is based on the idea of training multiple classifiers (primary) on the same sample, and the combination of its predictions conforms to some rules for new testing objects. Thus making it possible to collectively obtain more complex models than each model separately [14].

III. RELATED WORK

This research [1] uses the Global Terrorist Database (GTD) to learn to forecast the perpetrators of terrorist attacks and provided data on the types of attacks, targets, and weapons in addition to location, year, and other attributes using the Random Forest, Decision Tree and Gradient Boosting methods. Research [6] Presents the best machine learning models that can be used on terrorism-related data to predict the most accurate terrorist groups responsible for attacks based on historical data in India by modeling the behavior of terrorist groups using machine learning algorithms such as J48, IBK, Naive Bayes and Ensemble Voting approaches. In research [27], Create a framework for terrorist attacks that predicts the use of the Global Terrorism Database (GTD). The research approach assumes that textual features may influence the enhanced ability of classifiers to predict the types of terrorist attacks. Fitur text is extracted and represented using text representation techniques such as Term Frequency-Inverse Document Frequency (TF-IDF), Bag of Words (Bow), and Word Embedding (W2vec), Extracted later combined with data set features. The results showed that combining textual features with key features improved prediction accuracy significantly in research [28] using hypothesis tests and regression models. From a practice perspective, exploring the characteristics identified in patterns can lead to prevention strategies, such as changes in the physical or systemic environment. On research

[29] presents new insights into groups and target intruders using data mining algorithms by proposing a framework using historical data to train machine learning classifiers and predict intruder groups and attack types based on selected features using J48 and IBK algorithms. Research [30] uses terrorist event predictions from the Global Terrorism Database (GTD) with support vector machine (SVM), naive Bayes(NB), and linear regression(LR) techniques. Two feature selection methods, including Minimal-redundancy maximal-relevancy (mRMR) and maximum relevance (Max-Relevance), are used to improve classification accuracy. On research [31], Determine whether members of different organized crime groups cooperate using intelligence from the Canadian province of Alberta, which centers on criminals and criminal groups involved in different types of crimes in multiple locations. Bayesian techniques are used to extract multilevel network analytical frameworks and random graph models to uncover determinants of criminal collaboration between groups.

IV. RESEARCH METHOD

Over the past 20 years, terrorism has become a critical influencing factor in international politics and is now marked by increased terrorist attacks across international borders [30]. The increase in cases of terror crimes in the country itself is of particular concern to institutions, especially officers, because terror crimes affect the country's stability and harm the community. Therefore, to know the list of priority scales of targets under investigation should be seen from the reasons factors that indicate the priority scale of the list of supervised members and impact the amount of security stability and comfort of the entire community.

In this study, priority scale indicator predictions were designed using discrete methods described to define the target field by studying its features to identify problems [28] using a supervised learning approach [5]. The agreements to be used are supervised classification learning, including Machine Learning Decision Tree, KNN, Naïve Bayes, and Bagging, and building a predictive model from which results are interpreted [6].

Based on Fig. 1, the first stage is to determine the background and formulation of the problem to be raised using the study of the research literature that has been done to validate the urgency of the problem raised. The next step is from the background and formulation of the above problems and then re-conducted literature studies to determine the purpose of the research and the scope of the research and deepen the sentiment analysis model that will be offered as a solution. The next step is to collect the data used from the intelligence source of the investigation process, which is a collection of datasets from each terror target located in the West Java region obtained from the data collection process taken from Raw Data CDR, Raw Data Medsos, Raw Data Surveillance, Raw Data Funding, and Raw Data BAP. The dataset consists of 1088 data with 21 attributes. Crime datasets have inherent geographic features where all data in the dataset is not distributed randomly [5]. To divide by data dimensions, we analyzed the dataset to find multiple attributes by selecting the most promising feature attributes [27], [32], potentially

contributing to identifying the perpetrators [1]. From the attribute-giving techniques in this study, datasets are divided between desire patterns and abilities.

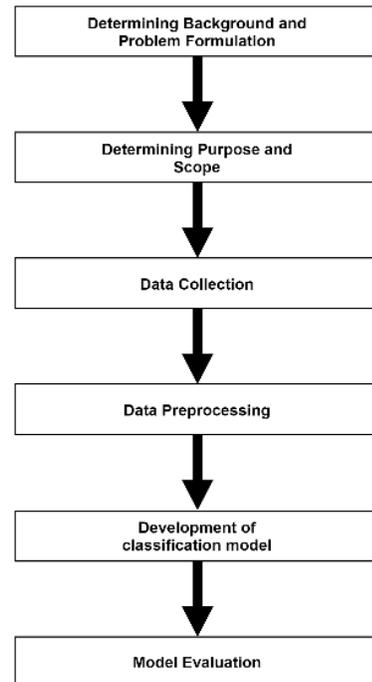


Fig. 1. The Flow of the Prediction System.

The desired stage pattern is based on the existence of intention and motivation, while the ability stage pattern consists of unique network patterns and patterns. It starts with the collection or retrieval of datasets, and then data is explored to determine inputs and outputs in dataset training and testing processes. The desire stage pattern is based on the existence of intention and motivation, while the ability stage pattern consists of unique patterns and network patterns. The attributes of these parameters are generated values from this priority scale which will be used as a result of research. The next step is to do data preparation which is done in several stages so that in the end, data can be used at the next stage. These stages include selecting and selecting data, transformation, and cleaning. The next step in building the classification will be through the stages that must be done in sequence and correctly. The steps are:

- Classification models are built for all three datasets using machine learning classification algorithms such as KNN, Naïve Bayes, and Decision Tree.
- Validation models are built for each algorithm to test base learner algorithms. The accuracy of the classifier can be further improved by using the Ensemble classifier.
- Validation models are built for each algorithm to test algorithms using Ensemble classification.
- The feature selection process aims to select subsets that can optically characterize the original data [30].

- In this study, classification algorithms were used to apply an ensemble approach and the use of attributes based on the best results of the Feature Selection process.
- Then the performance of the ensemble method is compared to the individual model.

After building a classification model, the stage that will be carried out in the evaluation stage. The stages are described below:

- Conducting a mode evaluation process using validation data tested with data testing then determines the value of Recall Rate, Precision Rate, and Accuracy Rate using confusion matrix measurement technique [27] where accuracy is single-handed, there is a significant difference between the number of Green, yellow and orange labels. Acquisition and accuracy can be used as criteria for classifier evaluation. This parameter is related to True Positive and False Positive (TP/FP), which refers to the number of positive predictions of right/wrong and True Negative and False Negative for the number of negative predictions of right/wrong (TN/FN). Confusion Matrix can measure machine learning performance in classification [11].
- Comparing the results with the other four methods (Decision Tree, KNN, Naïve Bayes, and Bagging).

A. Data Preparation

Data preparation will be done in several stages to obtain data used in the next stage eventually. These stages include Selection of data sampling, selection data, transformation, and cleaning with the priority scale sample dataset, and localized data system for pre-processed stages. Dataset used consists of 1088 Records with 23 Attributes. The primary purpose of the data is to find targets based on the risk of justice involvement in classification into three class groups, others referring to the green class, yellow class, and orange class. Then, perform pre-processed data and table determination. Separate data into training data and data testing 70:30 [7]. Use training data to train predictive models built on machine learning methods such as Decision Tree, Naïve Bayes, and K-Nearest Neighbour [19]. Finally, use data training to train validation models by performing k-fold cross validation processes [33] (k-fold =10) to get validation accuracy.

B. Model Development

Classification will go through the stages using machine learning classification algorithms such as KNN, Naïve Bayes, and Decision Tree. The Decision Tree Algorithm belongs to the family of trees used to generate decision trees. Naive Bayes is a probabilistic classifier belonging to the Bayes family. Furthermore, KNN is a lazy learning algorithm that implements the nearest K-neighbor algorithm. Validation models are built for each algorithm to test the base learner algorithm. The accuracy of the classifier can be further improved by using ensemble classifiers. Validation models are built for each algorithm to test the algorithm using Ensemble classification. The use of feature selection process because the feature aims to select a subset that can optically characterize

the original data [30]. Apply the ensemble approach and attributes based on the best results of the Feature Selection process. The three classification algorithms, KNN, Naïve Bayes, and Decision Tree, were given as inputs because the algorithm results had been analyzed earlier. Then the ensemble method performance was compared to the individual model.

Acquisition and accuracy can be used as criteria for classification evaluation. This parameter is associated with True Positive and False Positive (TP/FP), which refers to the number of true positive and false positive predictions, and True Negative and False Negative for true/false-negative predictions (TN/FN). Confusion Matrix can measure machine learning performance in classification [11]. Then the validation process is done using feature selection [33] to find which attributes can be used and which are not used by testing their accuracy. Furthermore, the determination of predictive accuracy values is based on the results of the cross-validation process. The results of the Feature Selection process determine the eight methods to be used based on the score used as which input can achieve the highest accuracy value and then compare the results with four other methods (Decision Tree, Naïve Bayes, K-Nearest Neighbour, and Bagging). Then the study concentrates on comparing predictive results to test data based on the results of feature training and then testing with data testing into the form of graphs or comparison tables.

The parameters and values of this priority scale results are then used as a dataset. Table I broadly describes the data obtained.

C. Confusion Matrix

For evaluation, we used the Confusion Matrix, as shown in Table II, to measure the accuracy of the classifier by calculating the ratio between the correctly predicted result and the number of samples. In this study, we will measure the level of accuracy, precision, and recall.

The explanations in Table II are:

- True Positive (TP) is the sum of one TRUE class that can be correctly predicted in the TRUE class.
- True Negative (TN) is the number of one FALSE class that can be predicted correctly in the FALSE class.
- False Positive (FP) is a condition where the TRUE class whose prediction is wrong in the FALSE class, while.
- False Negative (FN) is where the conditions in the FALSE class are predicted incorrectly in the TRUE class.
- The standard formula for calculating the degree of accuracy, precision, and recall is based on the confusion matrix as shown in Equations 1-3.

$$\text{Accuracy Rate} = \frac{(TP+TN)}{TP+TN+FP+FN} \quad (5)$$

$$\text{Precision Rate } (p) = \frac{TP}{(TP+FP)} \quad (6)$$

$$\text{Recall Rate } (r) = \frac{TP}{(TP+FN)} \quad (7)$$

TABLE I. RESEARCH DATASET

Type Variable	Variable	Description
Attribute	Mobile phone password	There is a password (No password=0, There is a password=1)
	Mobile phone encrypted	Mobile phone starts encrypted (Unencrypted=0, Encrypted=1)
	Mobile phone off/on	The mobile phone starts to die (Mobile phone off=0, Mobile on=1)
	Radical site access	Access radical sites (Not accessing radical sites =0, accessing radical sites =1)
	The meeting is getting more intense	Frequent meetings (Not attending meetings =0, Joining meetings =1)
	Get away from the network	Remove from the network / lone wolf (Not removing = 0, Removing yourself = 1)
	Counter Surveillance (SV)	Under surveillance (Unsupervised=0, Supervised =1)
	Purchase of Materials and Weapons	Making purchases of materials (Not Making illegal purchases =0, Buying illegal goods=1)
	Visiting the prison	Visiting the prison (Not visiting the prison =0, visiting the prison =1)
	Passport creation	Make a passport (Not make a passport =0, Create a passport =1)
	Withdrawal of large amounts of funds	Withdrawing vast amounts of funds (Not withdrawing funds =0, Withdrawing funds=1)
	Have the essential ability to make bombs	Has a bombing base (Has no base=0, Has a base=1)
	Personal funding capabilities	Have a permanent or non-permanent job (Unemployment=0, Work=1)
	Active Target Network	Active activities (Inactive=0, Active=1)
	Active Network Training activities	Frequent training activities (Not taking training =0, Taking training =1)
	Network Funding Capabilities	Frequently collecting network funds (Not raising funds=0, Raising funds =1)
	Permission from leader	Frequent visits to the leader of the organization (Never visited = 0, Visited = 1)
	Status background	Have a personal status background (Has no background=0, has background=1)
	Family background	Having the involvement of family members as network actors (Not having =0, Having=1)
	Label	Green
Orange		If there is a desire and no ability
Yellow		If there is no desire and there is the ability
ID	Initials Name	The name of each individual

TABLE II. CONFUSION MATRIX

Correct Classification	Classified as	
	Positive (+)	Negative (-)
Positive (+)	True Positive (TP)	False Negative (FN)
Negative (-)	False Positive (FP)	True Negative (TN)

V. RESULT AND DISCUSSION

In this study, we researchers will explain and discuss the study's results following the methods discussed in the previous chapter. The flow in this chapter will be presented in the flowchart:

Based on Fig. 2, an analysis of business needs will be carried out so that the data mining built can meet the needs of the company's goals. Unit XYZ is also unable to determine with certainty and estimate each target managed in developing the investigation process. In practice, it is often difficult to

determine the priority scale of each target. To ensure that priority scales can run effectively and efficiently, a strategy that considers the appropriate priority scale is needed to support the acceleration of handling the monitored targets. In this study, we used a dataset based on the investigation process taken from 2020 to 2021 for an investigation with Multivariate characteristics, with characteristic Attributes Polynomial and Integer, consisting of 3 Classes, 1088 Records, and 23 Attributes. Then the process using Retrieve Operator will upload an Object in the form of sample data from the Repository. After that, adding a Subprocess, this Operator that will combine other operators for the Preprocess stage will handle the Attributes of the data sample that the Retrieve Operator has loaded. The pre-processing stage is to convert nominal to numeric, normalize and replace missing values. Because there are still incorrect Attributes in writing, the MAP function maps a specific value of the selected attribute to be changed to a new value.

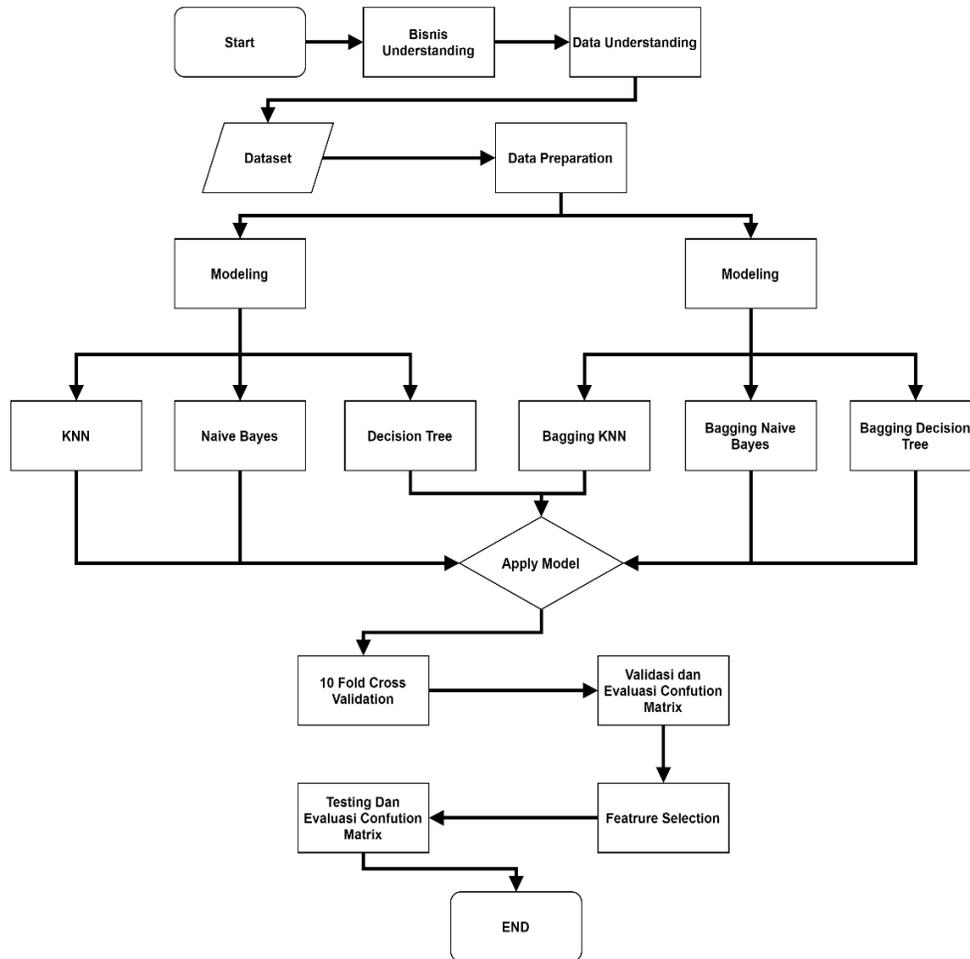


Fig. 2. Flowchart Result and Discussion.

Once the MAP process is complete, the next stage is Select Attributes. This Operator provides different filters to facilitate Attribute selection by selecting a subset of Attributes from Example Set and removing others. Only selected Attributes are sent to the output port. The rest was removed from Example Set. The next stage performs a Role Set to convert regular Attributes into Labels. This model calculation can be done by applying Set Role to turn the Priority Scale into a label. The next stage is to do the Split Data process based on the error score results. From this process, we are splitting the data 70:30. Next, separate the training data used to perform the validation test and the test data used to perform the test.

The Validation Test process tests several models using the Cross-Validation method. The accuracy of each model can be compared. The models tested were KNN, Naïve Bayes, and Decision Tree. The same Ensemble Bagging Model also uses three models to be tested: Bagging KNN, Bagging Naïve Bayes, and Bagging Decision Tree. The six models were chosen because they can process data available on RapidMiner. The validation testing process uses some training data. Validation accuracy results can be seen in Table III.

The next stage is the tuning process to get the best parameter value for feature selection and data mining. The feature selection results are carried out on the training data,

consisting of 21 input variables, using all the Attributes of the Dataset. The Feature Selection process uses the Forward Selection, Backward Elimination, and Optimize Selection parameters and produces two results from the Feature Selection process. The result of the first process is the feature selection process using single learner, KNN, Naïve Bayes, and Decision Tree. While the following result is the results of Feature Selection using the Bagging model that uses the Bagging KNN model, Bagging Naïve Bayes, and Bagging Decision Tree. Feature Selection results show that the Single Learner model obtained the highest result in the Decision Tree model using the Forward Selection-Decision Tree, getting 73.24% with 6 Attributes.

TABLE III. VALIDATION ACCURACY

Validation Model	Accuracy
KNN	71.26%
Naïve Bayes	65.88%
Decision Tree	70.35%
Bagging Decision Tree	69.68%
Bagging Naïve Bayes	66.01%
Bagging KNN	71.65%

Furthermore, the highest results in the Naïve Bayes model using Backward Elimination-Naïve Bayes get 71.79% with 13 Attributes. Moreover, lastly, the highest results obtained in the KNN model using Backward Elimination-KNN get 72.45% with 19 Attributes. The results of the Feature Selection get the result that bagging models obtained the highest results in bagging decision tree models using Forward Selection Bagging-Decision Tree get results 73.89% with 6 Attributes. Furthermore, the highest results were obtained in the Bagging Naïve Bayes model using Optimize Selection Bagging-Naïve Bayes got a result of 67.08% with 9 Attributes. Moreover, lastly, the highest results obtained in the Bagging KNN model using Backward Elimination Bagging KNN get 72.30% with 20 Attributes.

As shown in Table IV, the tuning process of the Single Learner model then produces the three best performance data, as seen below:

As shown in Table V, tuning the Bagging model produces the highest accuracy results for each model judging from the performance data for the following process. From the results of accuracy taken the best of each model as seen in Table V.

The processes presented in Fig. 3 are to conduct an Accuracy Testing process with Attributes based on the results of the Tuning process. The testing process uses models such as images with tuning parameters. The dataset used as input are divided into two separate parts.

The training process uses a dataset of 762 data with attributes that have been filtered according to the tuning process. The dataset used to produce the expected model is a dataset with 326 data and Attributes according to the tuning results in the testing process. In the testing model, there are two inputs, namely mods derived from the output of the training data model and Unl in the Apply Model, which comes from the output of the filter dataset and normalization. Accuracy Testing results have two results; the first is the process of Accuracy Testing using a single learner model that uses the Backward Elimination-KNN model using 19 Attributes, Backward Elimination-Naïve Bayes model uses 13 Attributes. The Forward Selection-Decision Tree model uses 6 Attributes. The following result is accuracy testing Ensemble Classifier using Bagging model that uses Bagging-Backward Elimination-KNN model using 20 Attributes, Bagging-Optimize Selection-Naïve Bayes model uses 9 Attributes, and Bagging-Forward Selection-Decision Tree model uses 6 Attributes. The Accuracy Testing results get results that for the Single Learner model obtained results for the KNN model using Backward Elimination get a result of 69.63%, with 19 Attributes.

Next is the Naïve Bayes model using Backward Elimination which produces 39.88% with 13 Attributes. Moreover, lastly, the Decision Tree model using Forward Selection gets 65.95% with 6 Attributes. The Accuracy Testing results get results that the Ensemble-Bagging model obtained results for the Bagging-KNN model using Backward Elimination get a result of 70.25% with 20 Attributes. Furthermore, the Naïve Bayes model using Optimize Selection results in 39.88% with 9 Attributes. Moreover, lastly, the

Decision Tree model using Forward Selection gets 68.10% with 6 Attributes.

The study used a data testing process tested using Feature Selection tuning validation data to see the accuracy of models in each class. By calculating the accuracy of some test data, the classification effectiveness can be known. This study uses 762 training data and 326 test data. The analysis compared and discussed three of the highest and different general classification performances, namely Backward Elimination K-Nearest Neighbor (KNN), Forward Selection Decision Tree (DT), and Elimination Naive Bayes (NB), as shown in Table VI and Table VII of the highest and other general classification Bagging performance bagging Backward Elimination K-Nearest Neighbor (KNN), Bagging Optimize Selection Naïve Bayes (NB), and Bagging Forward Selection Decision Tree (DT).

TABLE IV. SUMMARY FEATURE SELECTION SINGLE LEARNER

Backward Elimination			
	KNN	Naïve Bayes	Decision Tree
accuracy	72.45%	71.79%	71.80%
Forward Selection			
	KNN	Naïve Bayes	Decision Tree
accuracy	72.44%	66.80%	73.24%
Optimize Selection			
	KNN	Naïve Bayes	Decision Tree
accuracy	71.51%	66.80%	72.56%

TABLE V. SUMMARY FEATURE SELECTION BAGGING MODEL

Backward Elimination Bagging			
	KNN	Naïve Bayes	Decision Tree
accuracy	72.30%	66.93%	72.06%
Forward Selection Bagging			
	KNN	Naïve Bayes	Decision Tree
accuracy	72.31%	66.67%	73.89%
Optimize Selection Bagging			
	KNN	Naïve Bayes	Decision Tree
accuracy	71.78%	67.08%	73.76%

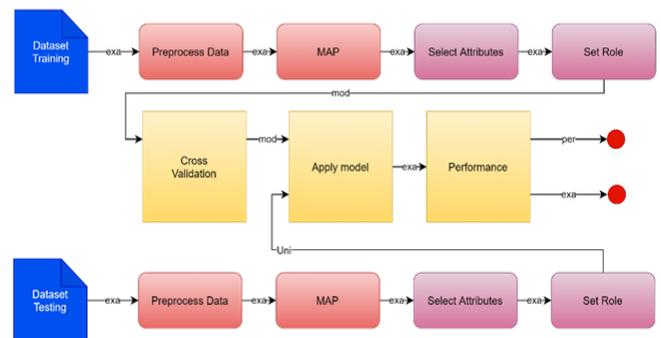


Fig. 3. Process Testing Flow.

TABLE VI. COMPARISON OF SINGLE LEARNER PREDICTION RESULTS

Predictive Actual	Backward Elimination KNN			Backward Elimination NB			Forward Selection DT		
	Green	Orange	Yellow	Green	Orange	Yellow	Green	Orange	Yellow
Green	1	0	0	0	0	0	1	0	0
Orange	0	60	27	1	130	193	0	66	45
Yellow	1	71	166	1	1	0	1	65	148
Total	2	131	193	2	131	193	2	131	193

TABLE VII. COMPARISON OF ENSEMBLE BAGGING PREDICTION RESULTS

Predictive Actual	Bagging Backward Elimination KNN			Bagging Optimize Selection NB			Bagging Forward Selection DT		
	Green	Orange	Yellow	Green	Orange	Yellow	Green	Orange	Yellow
Green	1	0	0	0	0	0	1	0	2
Orange	0	58	23	1	130	193	0	51	21
Yellow	1	73	170	1	1	0	1	80	170
Total	2	131	193	2	131	193	2	131	193

The first prediction modeling was done according to Table VI. It uses the Backward Elimination K-Nearest Neighbor method to predict an accuracy rate of 0.6963, the precision rate is 0.7957, and the recall rate is 0.6060. The following prediction model uses the Naïve Bayes Backward Elimination method and predicted accuracy of 0.3988; the precision rate was 0.1337, and the recall rate of prediction was 0.3308. The following prediction model used the Forward Selection Decision Tree method and obtained a prediction accuracy of 0.6595, the precision rate is 0.7621, and the recall rate of the prediction is 0.5902. For modeling, the second prediction is made according to Table VII. They are bagging the Backward Elimination K-Nearest Neighbor method that predicts an accuracy rate of 0.7025, the precision rate is 0.8043, and the recall rate is 0.6079. The following prediction model using the Bagging-Optimize Selection method Naïve Bayes obtained a prediction accuracy of 0.3988; the precision rate was 0.1337, and the recall rate of prediction was 0.3308. The following prediction model used the Bagging-Forward Selection Decision Tree method and obtained a prediction accuracy of 0.6810, the precision rate is 0.5730, and the recall rate of the prediction is 0.5900.

Of the several models tested, it is known that the one with the highest accuracy is Bagging-Backward Elimination-KNN, as seen in Table VIII.

Ensemble Bagging Classification Method to try to get better predictive accuracy. In this research, it is necessary to build a predictive model using Ensemble Bagging methods such as machine learning meta-algorithms to improve classification in terms of the stability and accuracy of classifications. It also reduces variance and helps avoid overfitting. The results were higher for Ensemble Bagging testing than testing using the single learner model. Of the several models tested, it is known that the one with the highest accuracy is Bagging-Backward Elimination-KNN, as seen in Fig. 4.

TABLE VIII. EVALUATION OF PREDICTION RESULTS

Classification Model	Accuracy	Precision	Recall
Backward Elimination-KNN	69.63%	79.57%	60.60%
Backward Elimination-Naïve Bayes	39.88%	13.37%	33.08%
Forward Selection-Decision Tree	65.95%	76.21%	59.02%
Bagging-Forward Selection-Decision Tree	68.10%	57.30%	59.00%
Bagging-Optimize Selection-Naïve Bayes	39.88%	13.37%	33.08%
Bagging-Backward Elimination-KNN	70.25%	80.43%	60.79%

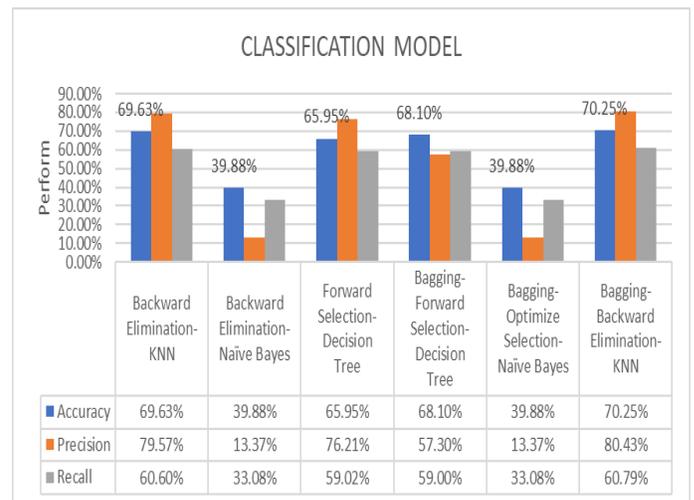


Fig. 4. Testing Performance.

VI. CONCLUSION

The experiments show how historical data or priority scale calculation patterns can be learned through data mining and generate new knowledge to predict future possibilities more accurately with other methods because the role of the priority scale is a picture of the consistency of individual targets in a behavior. The research shows that Bagging-KNN using Backward Elimination can be used, and the accuracy is 70.25%. Attributes that are considered to be no effect of Bagging KNN using Backward Elimination only one is There Is a Password. Influential attributes in determining the priority scale are Access to radical sites, Visiting prisons, Counter SV, Mobile Off, Permission from Amir, Active target network, Active Network training activities, Groups, Network Funding Capabilities, Personal funding capabilities, Family background, Background status, Removing from the network, Having basic bomb-making capabilities, Starting encrypted, Purchasing Materials and weapons, Making passports, Withdrawals in large numbers, increasingly intense Meetings and Territories. This research has discussed how historical data or priority scale calculation patterns can be learned through data mining and generate new knowledge to predict future possibilities more accurately with other methods. This study also compares the classification model of the previous ensemble research [6]. The Ensemble model is shown to improve the classification model with the research earlier model but uses an intelligence dataset instead of a dataset derived from GTD and uses the Ensemble method from previous studies using the J48, Naïve Bayes, IBK, and ensemble models using VOTE using the same model, only the VOTE Ensemble model use the same model. They are replaced using the Ensemble Bagging model. After implementing the research, it is known that the Bagging-KNN Ensemble Model using Backward Elimination can be used, and the accuracy reaches 70.25%. These results showed that the Ensemble model carried out against an existing model, Ensemble bagging KNN Backward Elimination, can increase the accuracy value by 0.62%.

Data mining with classification methods can predict the Priority Scale of each terrorist target. Using these prediction results, analysts can carry out an effective priority scale process in handling terrorism crimes so that the increase in cases of dissociation can be prevented and handled earlier. There are still many shortcomings in this study. For further work, we recommend increasing the number of variations of correlative features and large datasets so that it will help to improve the better performance of this study, namely external assessment features. In addition, more research is needed on feature selection grid methods so that each feature is more significant and very optimal for use in classification modeling.

REFERENCES

- [1] M. ALfatih, C. Li, and N. E. Saadalla, "Prediction of Groups Responsible for Terrorism Attack Using Tree Based Models," in Proceedings of the 2019 International Conference on Artificial Intelligence and Computer Science, Wuhan Hubei China, Jul. 2019, pp. 320–324.
- [2] H. A. H. Al-Sukhni, M. M. Saudi, and A. Ahmad, "A Review of Web Classifier Approach with Possible Research Direction to Detect Islamic Terrorists," International Journal of Advanced Computer Science and Applications, vol. 9, no. 12, p. 8, 2018.
- [3] M. Lim, A. Abdullah, and N. Jhanjhi, "Performance optimization of criminal network hidden link prediction model with deep reinforcement learning," Journal of King Saud University - Computer and Information Sciences, p. S1319157819308584, Jul. 2019.
- [4] S. V. Nath and S. Nath, "Crime Pattern Detection Using Data Mining," p. 4.
- [5] M. Farsi, A. Daneshkhah, A. H. Far, O. Chatrabgoun, and R. Montasari, "Crime Data Mining, Threat Analysis and Prediction," in Cyber Criminology, H. Jahankhani, Ed. Cham: Springer International Publishing, 2018, pp. 183–202.
- [6] V. T. Gundabathula, Department of Computer Science, Christ University, Bengaluru - 560029, Karnataka, India, V. Vaidhehi, and Department of Computer Science, Christ University, Bengaluru - 560029, Karnataka, India, "An Efficient Modelling of Terrorist Groups in India Using Machine Learning Algorithms," Indian Journal of Science and Technology, vol. 11, no. 15, pp. 1–10, Apr. 2018.
- [7] N. Hutagaol and S. Suharjito, "Predictive Modelling of Student Dropout Using Ensemble Classifier Method in Higher Education," Adv. sci. technol. eng. syst. j., vol. 4, no. 4, pp. 206–211, 2019.
- [8] P. O. Oketch, "An evaluation of hybrid machine learning classifier models for identification of terrorist groups in the aftermath of an attack," p. 146.
- [9] R. Sathyaraj and S. Prabu, "A hybrid approach to improve the quality of software fault prediction using Naïve Bayes and k-NN classification algorithm with ensemble method," p. 14.
- [10] I. Zafar, I. Y. Wuni, G. Q. P. Shen, S. Ahmed, and T. Yousaf, "A fuzzy synthetic evaluation analysis of time overrun risk factors in highway projects of terrorism-affected countries: the case of Pakistan," International Journal of Construction Management, pp. 1–19, Aug. 2019.
- [11] M. A. Jabbar and S. Suharjito, "Fraud Detection Call Detail Record Using Machine Learning in Telecommunications Company," Adv. sci. technol. eng. syst. j., vol. 5, no. 4, pp. 63–69, Jul. 2020.
- [12] F. Yuan, L. Lu, and Q. Zou, "Analysis of gene expression profiles of lung cancer subtypes with machine learning algorithms," Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease, vol. 1866, no. 8, p. 165822, Aug. 2020.
- [13] S. R. Bandekar and C. Vijayalakshmi, "Design and Analysis of Machine Learning Algorithms for the reduction of crime rates in India," Procedia Computer Science, vol. 172, pp. 122–127, 2020.
- [14] I. Lavrov and J. Domashova, "Constructor of compositions of machine learning models for solving classification problems," Procedia Computer Science, vol. 169, pp. 780–786, 2020.
- [15] D. Gibert, C. Mateu, and J. Planes, "HYDRA: A multimodal deep learning framework for malware classification," Computers & Security, vol. 95, p. 101873, Aug. 2020.
- [16] R. Victoriano, A. Paez, and J.-A. Carrasco, "Time, space, money, and social interaction: Using machine learning to classify people's mobility strategies through four key dimensions," Travel Behaviour and Society, vol. 20, pp. 1–11, Jul. 2020.
- [17] M. M. Mastoli, "Machine Learning Classification Algorithms for Predictive Analysis in Healthcare," vol. 06, no. 12, p. 5, 2019.
- [18] C. Frank, A. Habach, R. Seetan, and A. Wahbeh, "Predicting Smoking Status Using Machine Learning Algorithms and Statistical Analysis," Adv. sci. technol. eng. syst. j., vol. 3, no. 2, pp. 184–189, Mar. 2018.
- [19] Y. C. Widiyono and S. M. Isa, "Utilization of Data Mining to Predict Non-Performing Loan," Adv. sci. technol. eng. syst. j., vol. 5, no. 4, pp. 252–256, 2020.
- [20] F. Lopes, J. Agnelo, C. A. Teixeira, N. Laranjeiro, and J. Bernardino, "Automating orthogonal defect classification using machine learning algorithms," Future Generation Computer Systems, vol. 102, pp. 932–947, Jan. 2020.
- [21] G. N. Srivastava, A. S. Malwe, A. K. Sharma, V. Shastri, K. Hibare, and V. K. Sharma, "Molib: A machine learning based classification tool for the prediction of biofilm inhibitory molecules," Genomics, vol. 112, no. 4, pp. 2823–2832, Jul. 2020.
- [22] U. Pujiyanto, T. Widiyaningtyas, D. D. Prasetya, and B. Romadhon, "Penerapan algoritma naïve bayes classifier untuk klasifikasi judul

- skripsi dan tugas akhir berdasarkan Kelompok Bidang Keahlian," TEKNO, vol. 27, no. 1, p. 79, Jul. 2019.
- [23] R. Siva Subramanian and D. Prabha, "Customer behavior analysis using Naive Bayes with bagging homogeneous feature selection approach," *J Ambient Intell Human Comput*, vol. 12, no. 5, pp. 5105–5116, May 2021.
- [24] V. G. T. da Costa, S. M. Mastelini, A. C. P. de L. F. de Carvalho, and S. Barbon, "Making Data Stream Classification Tree-Based Ensembles Lighter," in *2018 7th Brazilian Conference on Intelligent Systems (BRACIS)*, Sao Paulo, Brazil, Oct. 2018, pp. 480–485.
- [25] S. Kabiraj, L. Akter, M. Raihan, N. J. Diba, E. Podder, and Md. M. Hassan, "Prediction of Recurrence and Non-recurrence Events of Breast Cancer using Bagging Algorithm," in *2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, Kharagpur, India, Jul. 2020, pp. 1–5.
- [26] R. O. Alabi et al., "Comparison of supervised machine learning classification techniques in prediction of locoregional recurrences in early oral tongue cancer," *International Journal of Medical Informatics*, vol. 136, p. 104068, Apr. 2020.
- [27] M. Abdalsalam, C. Li, A. Dahou, and S. Noor, "A Study of the Effects of Textual Features on Prediction of Terrorism Attacks in GTD Dataset," vol. 29, no. 2, p. 16, 2021.
- [28] G. Singer and M. Golan, "Identification of subgroups of terror attacks with shared characteristics for the purpose of preventing mass-casualty attacks: a data-mining approach," *Crime Sci*, vol. 8, no. 1, p. 14, Dec. 2019.
- [29] S. Nazir, M. A. Ghazanfar, N. R. Aljohani, M. A. Azam, and J. S. Alowibdi, "Data analysis to uncover intruder attacks using data mining techniques," in *2017 5th International Conference on Information and Communication Technology (ICoICT)*, Melaka, Malaysia, May 2017, pp. 1–6.
- [30] H. Mo, X. Meng, J. Li, and S. Zhao, "Terrorist event prediction based on revealing data," in *2017 IEEE 2nd International Conference on Big Data Analysis (ICBDA)*, Beijing, China, Mar. 2017, pp. 239–244.
- [31] J. A. Coutinho, T. Diviák, D. Bright, and J. Koskinen, "Multilevel determinants of collaboration between organised criminal groups," *Social Networks*, vol. 63, pp. 56–69, Oct. 2020.
- [32] F. N. Yacoub, M. Mamdouh, K. F. Kassem, M. Torki, and M. S. Abougabal, "Towards Terrorist Groups Prediction in Middle East and North Africa," p. 7.
- [33] O. Somantri, S. Wiyono, and D. Dairoh, "K-Means Method for Optimization of Student Final Project Theme Classification Using Support Vector Machine (SVM)," *SJI*, vol. 3, no. 1, pp. 34–45, Jun. 2016.

Evaluating the Effectiveness and Usability of AR-based OSH Application: HazHunt

Ahmad A. Kamal¹, Syahrul N. Junaini², Abdul H. Hashim³

Centre of Pre-University Studies Universiti Malaysia Sarawak, Kota Samarahan, Sarawak, Malaysia¹

Faculty of Computer Sciences and Information Technology, Universiti Malaysia Sarawak, Kota Samarahan, Sarawak, Malaysia²

Faculty of Cognitive Sciences and Human Development, Universiti Malaysia Sarawak, Kota Samarahan, Sarawak, Malaysia³

Abstract—This study investigates the effectiveness and usability level of an augmented reality (AR) application called HazHunt to improve occupational safety and health (OSH) training. Previous research shows that AR has been growing in popularity as an innovative tool to enhance hazard identification courses. HazHunt, a marker-based AR app, was first developed using Vuforia software with OSH experts' guidance. Then, two online sessions of hazard identification course were conducted, where the experimental group's (EG) training was enhanced with the implementation of HazHunt. Analysis shows that the EG scores better (mean = 13.82, $s = 3.38$, $n = 22$) than the CG (mean = 13.41, $s = 2.15$, $n = 22$) in the post-quiz, but this difference is statistically non-significant, with $t(21) = 0.48$ and one-tail $p = 0.32$. Reduced Instructional Motivation Survey (RIMMS) shows that EG participants obtained higher confidence levels among the Attention, relevance, confidence, satisfaction (ARCS) factors in learning motivation. The System Usability Scale (SUS) score of HazHunt recorded the maximum count of 'Good' rating (mean = 78.41, $n = 8$). It is concluded that HazHunt has positive impacts on enhancing OSH training in terms of effectiveness and motivational impact. HazHunt also scored a high SUS score among the EG.

Keywords—OSH training; computer-aided training; online learning; marker-based AR; AR-based application; SUS score

I. INTRODUCTION

Augmented reality (AR) has been proven to play a vital role among the nine pillars of Industrial Revolution 4.0 (IR4) in enhancing occupational safety and health (OSH) activities [1]. Among the activities include rehabilitation [2] and the innovation of serious games for safety training [3]. Particularly, OSH training has evolved tremendously over the years, but most organizations still rely upon traditional way to impart the knowledge. Today, complex working infrastructure present more hazardous environment for the safety and well-being of workers. Thus, AR is a relevant tool that has the capacity to enhance conventional training, suitable for current trend.

The effectiveness of using AR in improving academic performances has been proven for many teaching-learning process [4]. AR have been one of the many popular technologies, used by various institutions as the attractive and interactive elements provide positive effects towards learning performances. However, the overall effectiveness of AR-based technology intervention for professional training was reported to have a small effect on the outcome [5]. Interestingly, AR improved the overall effectiveness of vocational training for a large size effect [6]. This indicates the potential of deploying

applications built with AR technology for OSH related training may produce better effects in achieving training outcomes as the first issue highlighted in this study.

Motivational impact is the second issue addressed when training is conducted conventionally [7]. Conventional training has seen a worse decline in providing sufficient motivational impact, especially in this pandemic era [8]. To overcome this, other IR4 technology such as virtual reality has been implemented for the sole purpose of boosting the motivation among learners [9]. In this study, the technological tool of interest, chosen to enhance motivation is AR [10], as AR is proven to be able to increase motivational impact as opposed to the conventional training delivery methodology [11].

Consequently, today's era perceived many information and communications technology (ICT) tools developed to be used in teaching-learning for various benefits, which requires the usability level of such tools to be properly examined [12], as the third issue for this study. In easy terms, usability level refers to the indicator on a specific user in a specific context, the ability to utilize the tool in achieving a goal effectively, efficiently, and satisfactorily. Simply developing a mobile app is insufficient to conclude its contribution towards the users in said fields, which is the reason to evaluate the usability level.

Therefore, based on the three issues elaborated, the following research questions are established:

- 1) What is the effect of implementing AR technology tool towards academic performances in OSH training?
- 2) What is the motivational impact of deploying the AR tool as a part of the OSH training?
- 3) What is the usability level of the developed AR tool?

II. BACKGROUND STUDY

Based on the research questions stated, there is a necessity to study several important areas. These important areas include the importance of conducting OSH training among organizational members in the organization, the effect as well as motivational impact of conducting training conventionally, and the relevant usage of AR as an innovative tool that had been reported to improvise the content delivery of OSH training in recent times.

A. Importance of OSH Training

In general, training is important because it is a precondition to ensure employees can perform their job effectively and

efficiently and, in the context of OSH, to ensure they can perform a job safely on top of being effective and efficient. To perform a job safely, the employees must equip themselves with sufficient, suitable OSH knowledge and skills, including the knowledge of hazards and how to identify them, the skills to assess risks related to the job, and the identification of suitable control measures reduce the risks.

Clarke & Flitcroft [13] studied the impact of training intervention over 24 months period on a sample of 10 companies and found that the interventions had a significant impact on company safety culture and productivity and even reduced workplace accidents by 22% on average. They also reported a significant increase in employee motivation and safety participation. The study demonstrated that training could be one of the important factors in creating a positive OSH culture within organizations.

In addition to the inherent characteristics of training that makes it important to have a safe workplace, the provision of training is also part of the organization's legal duty stipulated in many statutory laws around the world. For instance, in the United Kingdom, the Health and Safety at Work etc. Act (1974) [14] requires the employer to provide whatever training as is necessary to ensure, so far as is reasonably practicable, the health and safety at work of their employees. Iteration of similar requirements can also be found in the United States Occupational Safety and Health Act (1970) [15] and its Standards, Australia's Work Health & Safety Act (2011) [16], India's Occupational Safety, Health and Working Conditions Code (2020) [17], Canada Labour Code (1985) [18], Japanese Industrial Safety and Health Act (1972) [15], France's Code Du Travail (Labour Code, 2015) [19] and Germany's Arbeitsschutzgesetz (Occupational Health and Safety Act, 1996) [20].

The importance of OSH training is further compounded by the establishment of government-owned or government-linked OSH training centres such as OSHA Training Institute in the United States, OSH Training Centre of Hong Kong, The Occupational Safety Training Institute of Korea and Malaysian National Institute of Occupational Safety and Health, among others.

B. Effect and Motivation of Conventional OSH Training

Conventional OSH training commonly refers to a method of transferring OSH knowledge from the instructor (or trainer) to the training participants using a presentation-based lecture over one to several days' duration in a classroom setting. The variability of training delivery methods will depend on the syllabus, time allocation, the experience of instructors, the setting and location of the classroom and availability of training equipment and material. At its worst, conventional OSH training involves dry lectures coupled with PowerPoint printouts given as handouts and the use of unnecessarily long and unstimulating videos with content unrelated to the participant's learning needs [21]. Despite the many potential weaknesses and variability of conventional training, it remained popular as ever, as evidenced by the Association of Talent Development's State of The Industry report (2019) [22], citing 40% of an organization's hours were spent in a traditional classroom setting.

According to Casey, Turner, Hu & Bancroft [23], two important concepts that are relevant for OSH conventional training effectiveness are training engagement and training transfer. Training transfer is the application of learned skills, generalization to work scenarios and maintenance over time. A recent study by Aziz & Osman [24] highlighted that 98.3% of respondents used what they learned in training at their respective workplaces after training completion. The respondent of the study comprised of those who attended the Malaysian National Institute of Occupational Safety and Health (NIOSH) conventional OSH training courses.

On the other hand, training engagement involves the trainee's engagement with the training during its delivery. Safety training engagement, according to the authors, is the combination of optimal cognitive, emotional and behavioral activity that drives a trainee's motivation to learn. However, there is little research done to answer the question of how engagement can be increased and which elements of engagement are crucial within the context of different types of safety training [23].

C. AR as an Innovative Tool for OSH Training

The impact of AR technology has been studied in various industrial and educational settings, such as the real estate and construction industry [25]. While AR has been applied to several applications—OSH training is one novel application. For example, Kamal et al. found that the AR-based application has increased OSH training participants' active learning behavior, engagement, and interest [26]. It has also been established that AR can aid employees in identifying possible risks and correcting them before they cause actual damage. Sports, architecture, entertainment, and health training [27] have benefited from AR-based training. AR-based training is projected to grow in popularity and functionality greatly in the next years—aligned with the explosion of mobile technology [28].

In general, AR-based training provides several possible benefits and downsides. When developing or conducting a training programme, it is critical to consider these elements to ensure that everyone participating is optimally matched for the training opportunity. For example, Vignali et al. [29] developed Wearable Augmented Reality for Employee Safety in Manufacturing Systems (W-Artemys). They intended to provide new technology assistance for enhancing employee safety while performing machine operations using AR. Additionally, it is critical to assess the performance of AR-based training tools to ensure that workers retain and implement the material. If these goals are accomplished, AR-based training may be an effective method of equipping employees with the information and skills.

Immersive environments based on AR and VR have been effectively used to train personnel in many high-risk sectors [30]. Here AR may be a helpful tool for OSH training for various reasons. First, it can aid workers in comprehending complicated procedures. Second, it can serve as a visual depiction of potentially dangerous circumstances, reinforcing safety messaging. For example, Laciok et al. [31] designed a scenario for a work-related accident using the XVR software environment. Thirdly, it can assist employees in identifying

and resolving dangers prior to the occurrence of real damage. Fourth, AR may be utilized to deliver hands-on training tailored to specific industrial or safety needs.

However, the fundamental challenge is that applications of AR to OSH are still in their infancy [32]. For example, current AR-based training systems lack the scalability to support complicated training requirements. In addition, they may not give adequate augmented visuals to offer cognitive support or support for educational techniques. Verily, the influence of AR on OSH training is dependent on the way it is integrated into the curriculum. Aromaa et al. [33] stressed that there is a need for more interactive virtual and augmented OSH training material.

Therefore, there are a few possible ways to integrate AR into OSH training. This includes bringing a serious game and gamification approach to offer personalized learning interaction. For example, Holtkamp et al. [34] integrate AR and serious gaming to train workers wearing the right protective equipment and how to properly use ladders on their job site. In addition, more and more new AR-based instructional technique is required to motivate trainees and students alike to learn OHS [35].

However, assessing and evaluating the success of online training can be challenging. For example, some individuals may find AR-based OSH training to be tedious or ineffective. This may result in the loss of critical information or skills that employees require to do their jobs. If employees are disengaged with the training, they may be less likely to retain and implement the material. All these possible disadvantages may result in diminished learning and retention. To tackle these challenges, Tarallo et al. [36] created a mobile solution targeted at streamlining and speeding up the flow of training content concerning safety-related issues among safety managers, workers, and casual users.

These disadvantages may be offset by the benefits of AR-based training, such as its accessibility and ease of use [37]. In general, both online and enhanced OSH trainings are projected to gain popularity soon. The primary issue will be to ensure that these training alternatives are meaningful and available to workers in various scenarios and locales.

D. AR Enhanced OSH Training

Indeed, AR has the potential to improve OSH training outcomes by increasing trainee motivation. For example, Vigoroso et al. [38] prove that AR-based training games can boost training effectiveness. In their study, they proved that machine operators' abilities and safety understanding has also been enhanced. The learning process may be further sped up by incorporating learners. While AR has been implemented in a few sectors, it has yet to be implemented in OSH training.

Therefore, an enhanced OSH training is proposed by deploying an AR-based application called HazHunt. This enhanced training suggests the following hypotheses to be tested where the dependent variables are the effectiveness of implementing HazHunt in the enhanced OSH training, the impact of HazHunt towards learners' motivation, and the usability level of HazHunt application.

H₁: The implementation of HazHunt has a significant effect on trainees' learning effectiveness.

H₂: The enhanced training using HazHunt has a positive impact on the motivation of trainees.

H₃: The HazHunt app is perceived to have a high usability level by the trainees.

In conclusion, the following research objectives (RO) are established to test the hypotheses, then answer the research questions outlined, which are to:

RO1. Measure the effectiveness of deploying HazHunt on academic performances.

RO2. Evaluate the motivational impact of using HazHunt in the enhanced OSH training.

RO3. Measure the usability level of HazHunt.

III. METHODOLOGY

In overall, the aim of this study is to enhance the conventional OSH training by implementing AR technology. So, the AR-based application, HazHunt was firstly developed. Then, to achieve the three ROs, two sessions of short OSH training course was held. Department of Occupational Safety and Health (DOSH) officers were appointed as the trainers for the short courses. After the HazHunt prototype is developed, an advertisement targeting university staff and students is launched to find potential participants. Due to the pandemic, the courses had to be delivered online via Google Meet. Registered participants are separated into CG and EG. The first session is delivered conventionally to a control group (CG), followed by the second session to an experimental group (EG). HazHunt was deployed in the second session to EG. For the purpose of data collection, RO1 is measured with mechanical and chemical hazard identification post-quiz, RO2 is measured with Reduced Instructional Materials Motivation Survey (RIMMS), and RO3 is measured with System Usability Scale (SUS), developed by John Brooke [39].

A. Phase I: Development of HazHunt

HazHunt is a marker-based AR application. It is developed using Vuforia software, with the assistance of OSH expert. Fig. 1 shows the HazHunt main menu interface. HazHunt contains AR embedded hazard pictograms to be scanned with the app to access the elaborations and quizzes. Fig. 2 shows the pictogram buttons to access specific hazards. These pictograms are the AR markers to trigger the elaboration of each hazard type, consisting of a video and description. Fig. 3 is an example of HazHunt in action, used to trigger the "skull and crossbones" elaboration by scanning its pictogram.

B. Phase II: Short Courses for CG

In the first session, the participants selected as the CG attended the course by conventional training method (online) as in Fig. 4. By the end of the session, a question and answer (Q&A) slot was conducted as shown in Fig. 5. Later, participants answered the post-quiz and the RIMMS.



Fig. 1. HazHunt Main Menu.



Fig. 2. HazHunt Pictogram Buttons to Access the Specific Hazard.



Fig. 3. HazHunt: the Skull and Crossbones.



Fig. 4. The Presentation in the Short Course Session with the CG.

C. Phase III: HazHunt Implementation for EG

The group of participants selected as EG underwent the second session (online). The trainers involved were prepared with the knowledge and ability to use HazHunt. The usage of HazHunt was integrated into the training in delivering related content. Prior to the session, participants were provided with a

HazHunt apk file to be downloaded and installed on their smart devices. During the session, the trainers instructed the EG to scan the marker that appeared on the shared screen using HazHunt, as shown in Fig. 6. In addition to the learning materials, Fig. 7 shows the HazHunt quiz to enrich the training with AR contents. At the end of the session, the EG answered the post-quiz, the RIMMS and the SUS.

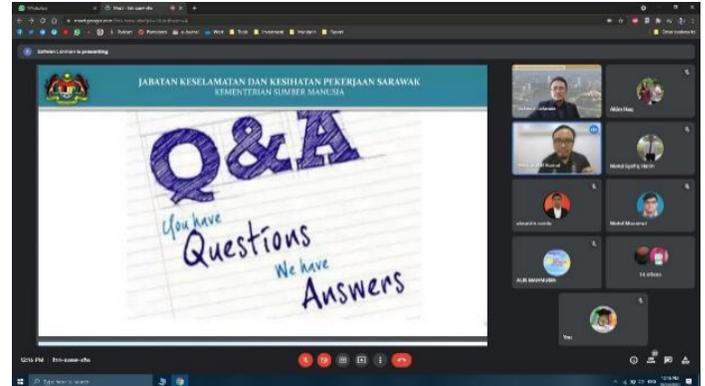


Fig. 5. Q&A Session in the Short Course Session with the CG.



Fig. 6. HazHunt was used in the Short Course with the EG.

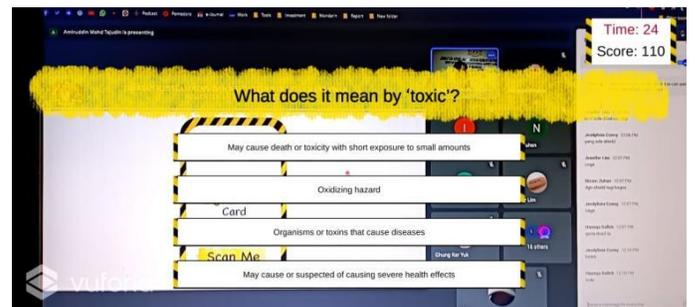


Fig. 7. The Quiz in HazHunt in the Short Course with the EG.

IV. RESULT

Courses were successfully conducted to both control groups (CG), and experimental groups (EG) involved 22 participants ($n = 22$) via Google Meet. Both courses were completed in October 2021.

A. Post-Quiz

The post-quiz consists of 18 questions derived from the topics covered in the course, which are corrosion, explosion, fire, general, health, oxidizer, and toxicity hazards. The post-quiz is analyzed to test the first hypothesis:

H₁: The implementation of HazHunt has a significant effect on trainees' learning effectiveness.

Table I shows the descriptive statistics of the post-quiz results from both groups. The post-quiz results of EG were hypothesized to be greater than CG's results. Based on the results, the EG has higher mean scores but also with higher standard deviation. This means the EG scores better than CG but vary larger among the participants, which shows that the HazHunt implementation does influence the learning effectiveness for the participants. This difference, however, was statistically non-significant, with $t(21) = 0.48$, $p = 0.32$ (one-tail), as can be observed from Table II.

B. RIMMS

The Reduced Instructional Motivation Survey (RIMMS) is a very useful tool to measure the learning motivations of the participants [40] to discern the differences between the CG and EG. The RIMMS is modified in conjunction with the training courses implementation. This survey is used to test the second hypothesis:

H₂: The enhanced training using HazHunt has a positive impact on the motivation of trainees.

Table III shows the descriptive statistics for CG, whereas Table IV shows the descriptive statistics for EG. The means from statements 1 to 3, statements 4 to 6, statements 7 to 9, and statements 10 to 12 contribute to the mean for attention, relevance, confidence, and satisfaction factors, respectively, as shown in Table V. Individually, the EG scores a higher mean in statements 2, 6, 7, and 11. However, EG participants only scored higher in the confidence factor in the ARCS model for learning motivation, collectively. Nonetheless, the enhanced training using HazHunt has a slightly higher positive impact on the participants' motivation to learn.

TABLE I. DESCRIPTIVE STATISTICS OF POST-QUIZ RESULTS

CG		EG	
Mean	13.41	Mean	13.82
Standard Deviation	2.15	Standard Deviation	3.38
N	22.00	N	22.00

TABLE II. THE T-TEST FOR THE TWO-SAMPLE ASSUMES UNEQUAL VARIANCES

Measures	EG	CG
Mean	13.82	13.41
N	22.00	22.00
Df	21.00	
t Stat	0.48	
p (T<=t) one-tail	0.32	
t Critical one-tail	1.69	

C. SUS Score

The SUS questionnaire is used to acquire the usability level [41] of HazHunt, distributed only to the participants in EG. The third hypothesis to be tested is:

H₃: The HazHunt app is perceived to have a high usability level by the trainees.

Table VI shows the descriptive statistics of the SUS scores for HazHunt. The scores range from 0 to 100, where HazHunt obtained a mean of 78.41 across 22 data collected, with the minimum and maximum scores being 62.50 and 95.00, respectively. According to the results shown in Table VII, the maximum scores with the count of 8 were recorded in the 'Good' rating, which means most of the scores were between 68 to 80.3 range. The second-highest scores count was recorded in 'Excellent' rating with the count of 10 participants to rate HazHunt above the 80.3 scores. None of the scores was equal to 68 under the 'Okay' and 'Poor' rating, but the 'Awful' rating consisted of 4 counts for the scores below 51. There might be several interpretations that can be perceived from these scores. However, factors non-related with HazHunt usability may have affected how participants provided the feedback in this questionnaire. Given that 81.8% (18 out of 22) participants provided above average for HazHunt SUS scores, HazHunt is perceived to have a high usability level.

TABLE III. DESCRIPTIVE STATISTICS OF RIMMS FOR CG

Statements	Mean	Standard Deviation
The quality of the slides helped to hold my attention.	4.58	0.50
The way the information is arranged on the slides helped keep my attention.	4.54	0.51
The variety of passages, exercises, illustrations, etc., helped keep my attention on the session.	4.42	0.58
It is clear to me how the contents of these slides are related to things I already know.	4.54	0.59
The content and style of writing in these slides convey the impression of me being able to apply the knowledge for my work.	4.58	0.50
The content of these user instructions will be useful to me.	4.50	0.51
As I sit more in the session, I was confident that I could learn well.	4.50	0.51
I was confident that I would be able to absorb all the knowledge and information.	4.42	0.50
The good organization of the slides helped me be confident that I would be able to learn.	4.42	0.58
I enjoyed learning with these slides so much that I was stimulated to keep on learning.	4.54	0.51
I really enjoyed learning through the slides.	4.54	0.51
It was a pleasure to learn with such a well-designed slide.	4.46	0.51

TABLE IV. DESCRIPTIVE STATISTICS OF RIMMS FOR CG

Statements	Mean	Standard Deviation
<i>The quality of the slides helped to hold my attention.</i>	4.53	0.51
<i>The way the information is arranged on the slides helped keep my attention.</i>	4.58	0.51
<i>The variety of passages, exercises, illustrations, etc., helped keep my attention on the session.</i>	4.42	0.69
<i>It is clear to me how the contents of these slides are related to things I already know.</i>	4.47	0.61
<i>The content and style of writing in these slides convey the impression of me being able to apply the knowledge for my work.</i>	4.53	0.51
<i>The content of these user instructions will be useful to me.</i>	4.63	0.50
<i>As I sit more in the session, I was confident that I could learn well.</i>	4.53	0.51
<i>I was confident that I would be able to absorb all the knowledge and information.</i>	4.47	0.51
<i>The good organization of the slides helped me be confident that I would be able to learn.</i>	4.47	0.51
<i>I enjoyed learning with these slides so much that I was stimulated to keep on learning.</i>	4.47	0.61
<i>I really enjoyed learning through the slides.</i>	4.58	0.51
<i>It was a pleasure to learn with such a well-designed slide.</i>	4.47	0.51

TABLE V. MEAN FOR ACRS FACTORS MODEL FOR CG AND EG.

Group	Attention	Relevance	Confidence	Satisfaction
CG	4.51	4.54	4.44	4.51
EG	4.51	4.54	4.49	4.51

TABLE VI. SUS SCORES FOR HAZHUNT IN EG

SUS SCORE	Value
Mean	78.41
Standard Deviation	10.02
Minimum	62.50
Maximum	95.00
Count	22.00

TABLE VII. SUS ADJECTIVAL RATING COUNT

SUS Adjectival Rating	Count
Poor	0
Awful	4
Okay	0
Good	10
Excellent	8

V. DISCUSSION

The HazHunt app is developed in the hope of enhancing the benefits of OSH training by applying AR-based technology. The analysis is done in the previous section shows positive

results in terms of participants' feedback, but the differences are not huge or significant between CG and EG. However, there may be several inferences that can be made from the analysis done as follows.

A. Online Training Environment

Online learning decreases a certain amount of enthusiasm among participants [42]. Furthermore, an online learning environment could have required the participants to master the usage of various hardware and software to get the most out of the training learning outcome [43]. Online training delivery is supposed to encourage active learning as the participants need to adapt to technology simultaneously looking at the course contents [44]. However, those who lack experience in virtual learning might face various technical issues before and during the training [45]. Additionally, the lack of interaction may have caused the participants to experience the feeling of being alienated or isolated [46]. These could have led the EG participant's HazHunt-enhanced training experience to be affected negatively.

B. Technical Inconveniences

Another common factor that may lead to bad feedback is the occurrence of the technical inconvenience with participants being alone with no physical assistance could be provided, which includes when handling AR technology [47]. In addition, throughout the online learning session, connectivity issues also occurred among participants, which could disrupt the AR demonstration [48], causing a bad impression of the technology and HazHunt. As of date, HazHunt is only available for Android users, and some device or operating software-specific bugs may hinder certain participant learning processes [49]. Although only a small percentage, some users among the participants might have viewed AR as an entertainment tool and rather perceived the technology lack the suitability to be used as a part of professional training [50].

C. Teaching Presence

For some people, teaching presence is required to ensure that the learning process happens as certain participants feel the need to have the trainer be physically closed to assist them to grow as the course progresses [51]. It is true that online training may replace face to face sessions in the cognitive area. Still, it could not replace meaningful interaction between trainer to participants and among peers [52], which caused the HazHunt enhanced training was not able to empower participants significantly.

VI. CONCLUSION

The purpose of this study was to determine the efficacy and usefulness of HazHunt, an AR-based application that was developed to enhance conventional OSH training. The marker-based AR app was built with the assistance of OSH expert using Vuforia. The enhancement of OSH training using HazHunt has been demonstrated to have a good influence on trainees' learning process. Based on the findings, AR technology has a positive effect towards academic performances and motivational impact of OSH training. Furthermore, the AR tool is shown to have a good usability level among the trainees, making the enhanced OSH training to have a new and meaningful experience.

ACKNOWLEDGMENT

This research was funded by the Universiti Malaysia Sarawak under the Cross-Disciplinary Research Grant scheme with the grant ID C09/CDRG/1837/2019.

REFERENCES

- [1] G. J. L. Micheli, G. Vitranò, and A. Calabrese, "Occupational Safety and Health Education and Training: A Latent Dirichlet Allocation Systematic Literature Review," in *Lecture Notes in Networks and Systems*, 2021, vol. 221 LNNS, pp. 491–502. doi: 10.1007/978-3-030-74608-7_61.
- [2] S. Chen, B. Hu, Y. Gao, Z. Liao, J. Li, and A. Hao, "Lower Limb Balance Rehabilitation of Post-stroke Patients Using an Evaluating and Training Combined Augmented Reality System," in *Adjunct Proceedings of the 2020 IEEE International Symposium on Mixed and Augmented Reality, ISMAR-Adjunct 2020*, 2020, pp. 217–218. doi: 10.1109/ISMAR-Adjunct51615.2020.00064.
- [3] L. Beever, S. Pop, and N. W. John, "Assisting Serious Games Level Design with an Augmented Reality Application and Workflow," in *Computer Graphics and Visual Computing, CGVC 2019*, 2019, pp. 9–17. doi: 10.2312/cgvc.20191253.
- [4] M. T. Jdaitawi and A. F. Kan'an, "A Decade of Research on the Effectiveness of Augmented Reality on Students with Special Disability in Higher Education," *Contemp. Educ. Technol.*, vol. 14, no. 1, pp. 1–16, 2022, doi: 10.30935/cedtech/11369.
- [5] X. Han, Y. Chen, Q. Feng, and H. Luo, "Augmented Reality in Professional Training: A Review of the Literature from 2001 to 2020," *Appl. Sci.*, vol. 12, no. 3, 2022, doi: 10.3390/app12031024.
- [6] F.-K. Chiang, X. Shang, and L. Qiao, "Augmented reality in vocational training: A systematic review of research and applications," *Comput. Human Behav.*, vol. 129, p. 107125, 2022, doi: <https://doi.org/10.1016/j.chb.2021.107125>.
- [7] G. Makransky, S. Borre-Gude, and R. E. Mayer, "Motivational and cognitive benefits of training in immersive virtual reality based on multiple assessments," *J. Comput. Assist. Learn.*, vol. 35, no. 6, pp. 691–707, 2019, doi: 10.1111/jcal.12375.
- [8] C. F. Yeap, N. Suhaimi, and M. K. M. Nasir, "Issues, Challenges, and Suggestions for Empowering Technical Vocational Education and Training Education during the COVID-19 Pandemic in Malaysia," *Creat. Educ.*, vol. 12, no. 08, pp. 1818–1839, 2021, doi: 10.4236/ce.2021.128138.
- [9] M. U. Sattar, S. Palaniappan, A. Lokman, A. Hassan, N. Shah, and Z. Riaz, "Effects of virtual reality training on medical students' learning motivation and competency," *Pakistan J. Med. Sci.*, vol. 35, no. 3, pp. 852–857, 2019, doi: 10.12669/pjms.35.3.44.
- [10] J. Bacca, S. Baldiris, R. Fabregat, and Kinshuk, "Framework for designing motivational augmented reality applications in vocational education and training," *Australas. J. Educ. Technol.*, vol. 35, no. 3, pp. 102–117, 2019, doi: 10.14742/ajet.4182.
- [11] J. Bacca, S. Baldiris, R. Fabregat, and Kinshuk, "Insights into the factors influencing student motivation in Augmented Reality learning experiences in Vocational Education and Training," *Front. Psychol.*, vol. 9, no. AUG, 2018, doi: 10.3389/fpsyg.2018.01486.
- [12] O. Alhadreti, "Assessing Academics' Perceptions of Blackboard Usability Using SUS and CSUQ: A Case Study during the COVID-19 Pandemic," *Int. J. Hum. Comput. Interact.*, vol. 37, no. 11, pp. 1003–1015, 2021, doi: 10.1080/10447318.2020.1861766.
- [13] S. Clarke and C. Flitcroft, "The effectiveness of training in promoting a positive OSH culture," *Eff. Train. Promot. a Posit. OSH Cult. Inst. Occup. Saf. & Heal. Wigst.*, 2013.
- [14] J. S. Humphreys, "Health and Safety at Work Act 1974: is it too late to teach an old dog new tricks?," *Policy Pract. Heal. Saf.*, vol. 5, no. 1, pp. 19–35, 2007.
- [15] L. Hornberger, "Occupational Safety and Health Act of 1970," *Clev. St. L. Rev.*, vol. 21, p. 1, 1972.
- [16] S. Australia, "Work Health and Safety Act 2012," *Aust. Fed. Gov.*, 2011.
- [17] A. Roychowdhury and K. Sarkar, "Labour reforms in a neo-liberal setting: Lessons from India," *Glob. Labour J.*, vol. 12, no. 1, 2021.
- [18] N. Solomon, "The Negotiation of First Agreements under the Canada Labour Code," *Relations Ind. Relations*, vol. 40, no. 3, pp. 458–472, 1985.
- [19] J. Dirringer, E. Dockès, E. Guillaume, P. Le Moal, and M. Marc, *Le Code du travail en sursis? Éditions Syllepse*, 2015.
- [20] A. Kelm, A. Meins-Becker, and M. Helmus, "Improving occupational health and safety by using advanced technologies and BIM," in *ISEC 2019 - 10th International Structural Engineering and Construction Conference*, 2019. doi: 10.14455/isec.res.2019.92.
- [21] H. E. Greene and C. L. Marcham, "Online vs. conventional safety training approaches," *Prof. Saf.*, vol. 64, no. 01, pp. 26–31, 2019.
- [22] M. Ho, "state of the industry: Talent development benchmarks and trends," *Alexandria, VA: ATD Research*, 2019.
- [23] T. Casey, N. Turner, X. Hu, and K. Bancroft, "Making safety training stickier: A richer model of safety training engagement and transfer," *J. Safety Res.*, vol. 78, pp. 303–313, 2021, doi: 10.1016/j.jsr.2021.06.004.
- [24] S. F. A. Aziz and F. Osman, "Does compulsory training improve occupational safety and health implementation? The case of Malaysian," *Saf. Sci.*, vol. 111, pp. 205–212, 2019.
- [25] A. Kelm, A. Meins-Becker, and M. Helmus, "Improving occupational health and safety by using advanced technologies and BIM," in *ISEC 2019 - 10th International Structural Engineering and Construction Conference*, 2019. doi: 10.14455/isec.res.2019.92.
- [26] A. A. Kamal, S. N. Junaini, A. H. Hashim, F. S. Sukor, and M. F. Said, "The Enhancement of OSH Training with an Augmented Reality-Based App," *Int. J. Online Biomed. Eng.*, vol. 17, no. 13, pp. 120–134, 2022, doi: 10.3991/ijoe.v17i13.24517.
- [27] A. R. Corvino, E. M. Garzillo, P. Arena, A. Cioffi, M. G. L. Monaco, and M. Lamberti, "Augmented Reality for Health and Safety Training Program Among Healthcare Workers: An Attempt at a Critical Review of the Literature," *Adv. Intell. Syst. Comput.*, vol. 876, pp. 711–715, 2019, doi: 10.1007/978-3-030-02053-8_108.
- [28] N. I. N. Ahmad and S. N. Junaini, "Augmented Reality for Learning Mathematics: A Systematic Literature Review," *Int. J. Emerg. Technol. Learn.*, vol. 15, no. 16, pp. 106–122, 2020, doi: 10.3991/ijet.v15i16.14961.
- [29] G. Vignali et al., "Development of a 4.0 industry application for increasing occupational safety: Guidelines for a correct approach," 2019. doi: 10.1109/ICE.2019.8792814.
- [30] S. Hasanazadeh, N. F. Polys, and J. M. De La Garza, "Presence, Mixed Reality, and Risk-Taking Behavior: A Study in Safety Interventions," *IEEE Trans. Vis. Comput. Graph.*, vol. 26, no. 5, pp. 2115–2125, 2020, doi: 10.1109/TVCG.2020.2973055.
- [31] V. Laciok, A. Bernatik, and M. Lesnak, "Experimental implementation of new technology into the area of teaching occupational safety for industry 4.0," *Int. J. Saf. Secur. Eng.*, vol. 10, no. 3, pp. 403–407, 2020, doi: 10.18280/ijssse.100313.
- [32] K. Gutsch and C. Droll, "Enabling or stressing? – Smart information use within industrial service operation," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 12199 LNCS, pp. 119–129, 2020, doi: 10.1007/978-3-030-49907-5_9.
- [33] S. Aromaa, A. Vääänen, I. Aaltonen, V. Goriachev, K. Helin, and J. Karjalainen, "Awareness of the real-world environment when using augmented reality head-mounted display," *Appl. Ergon.*, vol. 88, 2020, doi: 10.1016/j.apergo.2020.103145.
- [34] B. Holtkamp, M. Alshair, D. Biediger, M. Wilson, C. Yun, and K. Kim, "Enhancing subject matter assessments utilizing augmented reality and serious game techniques," in *ACM International Conference Proceeding Series*, 2019. doi: 10.1145/3337722.3337743.
- [35] A. Wahana and H. H. Marfuah, "The Use of Augmented Reality to Build Occupational Health and Safety (OHS) Learning Media," in *Journal of Physics: Conference Series*, 2021, vol. 1823, no. 1. doi: 10.1088/1742-6596/1823/1/012060.
- [36] A. Tarallo et al., "An augmented and interactive AID for occupational safety," in *30th European Safety and Reliability Conference, ESREL*

- 2020 and 15th Probabilistic Safety Assessment and Management Conference, PSAM 2020, 2020, pp. 1787–1791.
- [37] A. A. Kamal and S. N. Junaini, “The effects of design-based learning in teaching augmented reality for pre-university students in the ict competency course,” *Int. J. Sci. Technol. Res.*, vol. 8, no. 12, pp. 2726–2730, 2019.
- [38] L. Vigoroso, F. Caffaro, M. M. Cremasco, and E. Cavallo, “Innovating occupational safety training: A scoping review on digital games and possible applications in agriculture,” *Int. J. Environ. Res. Public Health*, vol. 18, no. 4, pp. 1–23, 2021, doi: 10.3390/ijerph18041868.
- [39] JohBrooken, “SUS : A Retrospective,” *J. Usability Stud.*, vol. 8, no. 2, pp. 29–40, 2013, [Online]. Available: http://www.usabilityprofessionals.org/upa%7B%5C_%7Dpublications/jus/2013february/brooke1.html%7B%5C%25%7D5Cnhttp://www.usability.gov/how-to-and-tools/methods/system-usability-scale.html.
- [40] F. Yang and Y. M. Goh, “VR and MR technology for safety management education: An authentic learning approach,” *Saf. Sci.*, vol. 148, p. 105645, 2022.
- [41] G. Rasheed, M. Khan, N. Malik, and A. Akhuzada, “Measuring learnability through virtual reality laboratory application: A user study,” *Sustain.*, vol. 13, no. 19, 2021, doi: 10.3390/su131910812.
- [42] M. N. I. Saleh, R. Sari, and P. Alim, “University Students’ Perception on The Implementation of Online Learning During The Covid-19,” *Nazhruna J. Pendidik. Islam*, vol. 4, no. 1, pp. 1–17, 2021, doi: 10.31538/nzh.v4i1.1022.
- [43] A. Yuliansyah and M. Ayu, “The implementation of project-based assignment in online learning during covid-19,” *J. English Lang. Teach. Learn.*, vol. 2, no. 1, pp. 32–38, 2021.
- [44] S. Sandrone, G. Scott, W. J. Anderson, and K. Musunuru, “Active learning-based STEM education for in-person and online learning,” *Cell*, vol. 184, no. 6, pp. 1409–1414, 2021, doi: 10.1016/j.cell.2021.01.045.
- [45] M. Bączek Michał and Zagańczyk-Bączek, M. Szpringer, A. Jaroszyński, and B. Wożakowska-Kapłon, “Students’ perception of online learning during the COVID-19 pandemic: A survey study of Polish medical students,” *Medicine (Baltimore)*, vol. 100, no. 7, p. e24821, 2021, doi: 10.1097/MD.00000000000024821.
- [46] L. H. Yang, “Online Learning Experiences of Irish University Students during the COVID-19 Pandemic,” *All Irel. J. High. Educ.*, vol. 13, no. 1, 2021.
- [47] A. Palanci and Z. Turan, “How Does the Use of the Augmented Reality Technology in Mathematics Education Affect Learning Processes?: A Systematic Review,” *Uluslararası E-İtım Programları ve Ö-İtım Çalışmaları Derg.*, vol. 11, no. 1, pp. 89–110, 2021.
- [48] I. Ü. Yapıcı and F. Karakoyun, “Using Augmented Reality in Biology Teaching,” *Malaysian Online J. Educ. Technol.*, vol. 9, no. 3, pp. 40–51, 2021.
- [49] G. Keçeci, P. Yildirim, and F. K. Zengin, “Opinions of Secondary School Students on the Use of Mobile Augmented Reality Technology in Science Teaching,” *J. Sci. Learn.*, vol. 4, no. 4, pp. 327–336, 2021.
- [50] P. Vijayakumar and A. Lawrence, “Virtual Reality--How Real Is the Indian Education Field?,” 2021.
- [51] J. Singh, K. Steele, and L. Singh, “Combining the Best of Online and Face-to-Face Learning: Hybrid and Blended Learning Approach for COVID-19, Post Vaccine, & Post-Pandemic World,” *J. Educ. Technol. Syst.*, vol. 50, no. 2, pp. 140–171, 2021, doi: 10.1177/00472395211047865.
- [52] A. Anggrawan and Q. S. Jihadil, “Comparative analysis of online E-learning and face to face learning: An experimental study,” 2018. doi: 10.1109/IAC.2018.8780495.

Efficacy of the Image Augmentation Method using CNN Transfer Learning in Identification of Timber Defect

Teo Hong Chun¹, Umami Raba'ah Hashim², Sabrina Ahmad³, Lizawati Salahuddin⁴, Ngo Hea Choon⁵, Kasturi Kanchymalay⁶

Department of Information Technology and Communication, Politeknik Mersing, Johor, Malaysia¹
Fakulti Teknologi Maklumat dan Komunikasi (FTMK), Universiti Teknikal Malaysia Melaka (UTeM), Melaka, Malaysia^{2, 3, 4, 5, 6}

Abstract—This paper discusses the efficacy of the data augmentation method deployed in many Convolutional Neural Network (CNN) algorithms for determining timber defect in four timber species from Malaysia. A sequence of morphological transformation, involving x-reflection and rotation, was executed in the timber defect augmentation dataset for aiding CNN model training and generating the finest CNN models which offer the best classification performance in determining timber defect. For further assessing the CNN algorithms' classification performance, several deep learning hyperparameters were tried on the Merbau timber species by utilising epoch as well as learning rate. A comparison of the classification performance was then done between other timber classes, namely KSK, Meranti, and Rubberwood. According to the results, the ResNet50 algorithm, which has its basis in the transfer learning methodology, outclasses other CNN algorithms (ShuffleNet, AlexNet, MobileNetV2, NASNetMobile, and GoogLeNet) with the best classification accuracy of 94.59% using the data augmentation method. Furthermore, the outcomes indicate that utilising an augmentation methodology not just addresses the issue of a limited dataset but also enhances CNN classification output by 5.78% with the support of T-test that demonstrates a significant difference across all CNN algorithms except for Alexnet. Our study on hyperparameter optimisation by utilising learning rate as well as epoch is sufficient to infer that a greater number of epoch and learning rate does not deliver superior precision in CNN classification. The experimental findings suggest that the proposed methods improved CNN algorithms classification performance in identification of timber defect while tackling the imbalanced and limited dataset challenges.

Keywords—Convolutional neural network; deep learning; defect identification; image augmentation; transfer learning

I. INTRODUCTION

Of late, a multiple integration of the artificial intelligence algorithm and image processing approach has been studied for determining timber defects as the image segmentation methodology single-handedly cannot precisely categorise such defects. Even though many machine learning algorithms have shown considerable recognition rates for different kinds of timber defects [1][2][3], the present manual feature extraction procedure executed in machine learning is considered quite taxing due to its vulnerability to multiple feature attributes within the distinctive appearance of the timber. Therefore, the convolutional neural network (CNN) is deployed for

addressing the complex procedure of feature extraction in machine learning as deep learning algorithms not just offer superior classification performance but also provide an automatic feature extraction process which is tailored to the imminent problem during the training procedure. CNN is a deep learning algorithm which blends hierarchical and multilayer network architectures. Its distinctive architecture allows the algorithm to mine diverse abstract representations on the basis of a designated level of features while aiding CNN to imbibe the complex matter with an improved feature set [4]. CNN has exhibited its competence by outclassing traditional computer algorithms when it comes to object identification and image-based classification. Even though such computer algorithms have been utilised for scrutinising actual images in various fields since the 1970s, these pre-fabricated texture algorithms have to be built based on image domain specificities, which is a major issue [5]. CNN demonstrated its ability when DenseNet was able to attain a classification precision of 98.75% while [6] assessing the performance of four novel CNN architectures with pre-determined texturing techniques in the timber domain. Jung et al. [7] made a comparison of the performance of three CNN architecture depths by categorising into four kinds of defect classes. The outcomes indicated that deep CNN attained the best classification precision of 99.8% in determining timber defect, albeit with greater computational time because of the deeper network architecture. Although the precision of the architectures varies a bit, this shows that they are both viable solutions to the issues of timber classification.

Thus, the deep learning methodology presents a good ability for data mining and knowledge breakthrough in the domain of timber defect detection. However, one of the significant challenges in execution of deep learning is data disparity. On one hand, raw data is now more and more accessible; on the other hand, most datasets have unbalanced distributions with some object classes exhibiting plentiful representation and others possessing inadequate representations, like timber defects. Data disparity in deep learning could trigger inadvertent errors with possibly substantial consequences, particularly in classification tasks in which the lopsided distribution of class instances compels classification algorithms to trigger inductive bias with regards to the majority class. This causes a substandard classification performance because of lesser detection of the minority

samples [8][9]. For dealing with the challenges associated with imbalanced datasets, data augmentation was generally deployed to produce supplementary samples as shown by [10] in their study on imbalanced toxic comments classification by utilising deep learning as well as data augmentation. Moreover, Hu et al. [11] noted that deep learning approaches are not much deployed in the timber industry because of the inadequate quantity of defect datasets necessary for CNN training. Other aspects contributing to the dearth of timber defect images are the outlay incurred for gathering such images and the rigorous manual labelling procedure. One of the effectual techniques for utilising CNNs on minor datasets is espousing transfer learning that encompasses dropping a pre-trained CNN's classifier layer and adjusting it for the target dataset [12]. Thus, the transfer learning and data augmentation approaches might be the solution for addressing class disparity and limited-data problems in timber defect datasets.

This study advocated the use of deep learning approaches, in this case a Convolutional Neural Network (CNN) algorithm to address the complex procedure of feature extraction in machine learning as deep learning algorithms itself not only offer superior classification performance but also provide an automatic feature extraction process which is tailored to the imminent problem during the training procedure. In order to utilize the advantage and capabilities of CNN, both transfer learning and data augmentation technique are used to cater for imbalanced and limited size timber defect dataset. The data augmentation technique employed in this study would involve implementation of various morphological transformation during the image pre-processing process to increase the diversity of timber defect dataset, while the proposed transfer learning method will be applied to several CNN algorithms (ShuffleNet, AlexNet, MobileNetV2, ResNet50, NASNetMobile, and GoogLeNet) in search of highest CNN classification performance across the timber species via multiple combinations of learning rate and epoch parameters. In addition, both transfer learning and data augmentation technique proposed in this research is necessary to avoid overfitting during the training of deep learning algorithm and achieve greater accuracy for timber defect identification. Besides studying the efficacy of these two methodologies, this research assesses the performance of several CNN algorithms across the four timber classes from Malaysia.

II. METHODOLOGY

A. Overview of Approach

This portion elaborates the research approach devised to assess transfer learning and data augmentation efficacy based on numerous CNN algorithms to determine timber defects for

four species in Malaysia. The initial timber pictures are labelled and classified using species and timber defect categories. The timber defect dataset provided by the Universiti Teknikal Melaka Malaysia (UTeM) is used for this research [13]. Meranti, Rubberwood, Kembang Semangkuk (KSK), and Merbau species were used for this study. Data augmentation approaches were applied to the images representing timber defects to assess process efficacy for CNN classification. Original and enhanced images were sized using the inputs corresponding to the CNN techniques. Subsequently, ShuffleNet, AlexNet, MobileNetV2, ResNet50, NASNetMobile, and GoogLeNet were used for testing the transfer learning approach. The techniques were further assessed to determine hyperparameter configurations for optimal CNN classification effectiveness of timber defects. Lastly, the enhanced dataset was tested for classification efficacy using several deep learning hyperparameter settings for different timber specimens; data were gathered and assessed.

B. Data Augmentation

Data augmentation is regarded as the standard approach in deep learning, specifically when there are fewer data samples. The study considered the Meranti, Merbau, KSK, and Rubberwood timber species; experimental specimens were created using 1600 images representing timber defects. The dataset comprises eight timber defect categories (brown stain, blue stain, knot, borer holes, rot, bark pocket, wane, and split), along with one set of clear timber specimens. The dataset was enhanced using the augmentation technique that allowed data maximisation by processing original images. Several morphological changes (x-reflection and rotation) and different orientations of the original images were used to augment the dataset and enhance CNN timber defect detection accuracy. Researchers [14] assert that several morphological changes also reduce overfitting challenges associated with deep learning by allowing additional variations of the original information. Defect images were represented using 10-degree rotation versions to comprehensively depict defects as they appear on timber surfaces based on the direction of feeding. Image enhancement and other pre-processing approaches helped enhance the original dataset to ten times its initial size, i.e., comprising 16000 timber defect images. CNN architecture training and testing were based on the original and enhanced datasets. Table I presents some information concerning the morphological changes implemented during the image addition pre-processing steps to enhance dataset diversity.

- Rotation range 10–350°; rotates every image by 10°
- X – Reflection with 0°, 90°, 180° and 270° rotation

TABLE I. AUGMENTED SAMPLES USING MULTIPLE MORPHOLOGICAL TRANSFORMATION

Rotation	Image	Rotation	Image	Rotation	Image	Rotation	Image
Original		10°		20°		30°	
70°		90°		130°		180°	
210°		270°		310°		330°	
X-Reflection		X-Reflection 90°		X-Reflection 180°		X-Reflection 270°	

C. Transfer Learning using Convolutional Neural Network (CNN) Algorithms

Using transfer learning for CNN is an effective and practical approach to train deep learning models during labelled specimen scarcity. Transfer learning offers the versatility to change initially-trained systems and tune them using domain-based information. A model trained using a broader dataset is used, and specific weights are preferred over initial model training. Transfer learning acts as a potent approach for enhancing learning speed for image categorization and identification jobs. Higher speeds are possible because previous training jobs are employed, and knowledge is reused to enhance learning speed for new or relatively difficult data models [15]. Further, [16] established that transfer learning was usable for VGG16, AlexNet, and ResNet152 to identify timber defects. ResNet provided 80.6% accuracy when contrasted against the speedier R-CNN framework and other previously-trained transfer learning systems. Moreover, transfer learning is a vital approach to reduce overfitting when training deep learning models [17]. For CNN systems, transfer learning is implemented by assigning convolutional layer weights equal to the starting values for fresh classification problems than the complete network comprising fully-connected layers. Moreover, this approach is specifically beneficial to address the difficulty concerning learning classifiers where strong performance is required, but training samples are limited [18].

- AlexNet is among the initial noteworthy CNN system used for the ImageNet dataset to classify objects [19]. The system consists of five and three convolutional and fully-connected layers, 500,000 neurons, and 58 million parameters. A SoftMax classifier is used after the fully-connected layers; it outputs the likelihood values for a relevant class [16].
- ShuffleNet works under computational capability constraints, and this deep learning architecture is tuned for mobile devices. The system comprises convolutional and maximum pooling layers, 3 ShuffleUnit elements, global pooling, and fully-connected layers [20]. The model comprises point-by-

point pairing convolution and shuffles channels to maximise classification effectiveness while controlling the need for higher computational abilities [21].

- GoogLeNet is a differentiated neural network framework implementing a novel organisational system called “Inception Module”. This module implements several convolution operations along with filter concatenation for subsequent layers. Overall, the system comprises 27 layers (pooling layers included). There are 9 inception elements having maxpooling and convolution processes [22].
- ResNet50 is short for residual network (ResNet) and it comprises 50 layers. Contrasting to other CNN algorithms that amass manifold convolutional layers within their architectures, the ResNet50 architecture comprises diverse sets of identical layers and ascertains blocks which are utilised for signifying the usage of prior layers in the network [23]. Even though the network architecture is quite deeper, the quantity of parameters is quite smaller as against other equivalents [24].
- NASNetMobile is a neural architecture search net (NASNet) variation which emphasises on the mobile and embedded platforms. The architecture’s central structure utilises data-led and intelligent methodologies to construct network frameworks which are optimised through reinforcement learning. It generates a feature map by deploying repeated operations on either convolutional cells (reduction cell and normal cell) throughout the architecture [25]. The architecture comprises 12 cells having 5.3 million parameters [26].
- The MobileNetV2 architecture is enhanced for mobile computing. It decreases consumption of memory and delivers speed at a lower cost while removing overfitting on minor datasets [27]. The inverted residual and depth wise separable convolution are two of the main aspects in the MobileNetV2 architecture that comprises 32 entirely convolutional filters and 19 residual bottleneck layers [28]. MobileNetV2 has 3.47 million parameters [26].

In this research, transfer learning is implemented by altering the output class of both fully connected layer and classification layer of the CNN algorithms according to the number of classes in the timber defect dataset (9 classes). However, the other important CNN layer such as convolutional layer, activation function (ReLU), pooling layer are kept in their original algorithm architecture. Furthermore, all six CNN algorithms were fine-tuned to match the data in this article by retraining the weightage of each CNN layer. For model fine-tuning, each of the CNN algorithm was trained 48 times using different combinations of learning rate and epoch parameters across multiple timber species. While other training options such as stochastic gradient descent with momentum (SGDM) optimizer and batch size (10) were used and maintained throughout the training. Even though a decent predictive performance entails a huge number of annotated datasets, transfer learning is frequently utilised to adjust for data paucity. As per the observations by [29], the limited dataset will be sufficient for the remaining layers for learning the features in the pertinent domains, since the architecture had attained vital features like corners in their initial few layers. Thus, transfer learning has been seen to be a mostly effective technique for training neural networks having a limited dataset, and offers a significant promise in the domain of classifying timber defects.

D. Hyperparameter Optimization

Hyperparameter optimisation is a primary constituent essential in deep learning training for enhancing the CNN algorithms' performance. Even though the procedure is quite tough as well as time-consuming, the fine-tuning is necessary to warrant the high classification performance of these algorithms since these are the variables which the model is unable to learn independently. While there exist many proposals on automatic optimisation methods, each has its own merits and demerits when implemented for diverse kinds of problems [30]. Batch size, learning rate, and training epochs are few of the hyperparameters which are adjusted as per the intricacy of the problems or datasets because the model has to possess adequate capability for prediction tasks while evading over-fitting [31]. The learning rate is a vital hyperparameter in deep learning since it outlines the step size at every iteration for the objective function to congregate [32]. The learning rate is augmented by a superior learning rate; however, the gradient might fluctuate around a local minimum value or perhaps fail to congregate. A minor learning rate would congregate smoothly but with a substantial rise in model training time because of supplementary training epochs. Notably, in case the gradient is trapped at local minima, visible progress is achieved at the expense of computational outlay [33]. With a proper rate of learning, the objective function has to be able to congregate to a global minimum within a decent period of time. Conversely, the number of epochs can be ascertained by the size of the training set and has to be adjusted by progressively raising its value till validation precision starts to fall, signifying model overfitting. The deep learning model typically congregates in a few epochs and the following epochs might drive supplementary execution time as well as overfitting. This can be evaded through an early halting approach. This approach is a kind of regularisation wherein the model training is halted beforehand when the validation precision does not

enhance following a specific number of successive epochs. To sum up, ascertaining the apt hyperparameters is vital for warranting the utmost performance of learning algorithms, thus creating a model timber defect identification setup in this study. We adjust the learning rate (0.001 and 0.0001) and the quantity of training epochs (50, 100 and 200) to ascertain the top CNN classification performance for determining timber defect.

III. RESULT AND DISCUSSION

In this paper, multiple CNN classification performance was examined via analysis of the concerning classification accuracy measures. With the accuracy signifying the measure pertaining to true defects versus the predicted defects, this study focuses on highlighting the classification performance pertaining to the put forward augmentation method via comparison versus those six CNN algorithms. Again, comparison was performed for the detailed classification performance with regards to the put forward feature versus four Malaysia timber species, namely Merbau, Meranti, Rubberwood and KSK. By employing both epochs and learning rate, multiple tuning pertaining to both hyperparameters were evaluated to identify the best CNN training optimisation and determine timber defect. Table II displays the classification accuracy with regards to various CNN algorithms across both non-augmented and augmented timber defect dataset along with hyperparameter tuning. While the classification accuracy pertaining to both MobileNetV2 and ResNet50 was seen to enhance considerably, ResNet50 was introduced to show a better performance at 0.01 learning rate and 100 epochs. Thus, this signifies that the highest accuracy rate is displayed by 94.59% classification rate from augmented Rubberwood dataset versus other timber species as well as CNN algorithms. The greatest effect was cast by the augmented dataset with synthetic data on the Rubberwood dataset, wherein accuracy enhanced by almost 10.37% from 82.96% to 93.33%, while Merbau dataset displayed the lowest impact, displaying reduced accuracy to 86.74% from 89.63%. The highest classification accuracy of 94.07% was achieved via GoogLeNet employing 0.001 learning rate and 200 epochs in Meranti dataset. Based on the tables, it can be seen that augmented Rubberwood dataset distinctly had the highest accuracy enhancement at 11.11% from 81.48% to 92.59% with 0.001 learning rate and 50 epochs. Even though Merbau is regarded to be the most ineffective augmented dataset, which reduced the classification accuracy to 75.85% from 84.44%, the overall classification accuracy pertaining to GoogLeNet algorithm encompassing four different types of timber species was seen to rise by 3.18%. With regards to AlexNet, the highest classification performance was achieved at 92.81% by employing the Rubberwood dataset that had hyperparameters of 0.0001 learning rate and 50 epochs. Using data augmentation was seen to enhance the classification accuracy of the algorithm by 22.87%, i.e., from 68.69% to 91.56%. However, the augmentation technique also comes along with adverse impact, wherein the augmented dataset pertaining to the KSK species made the algorithm to overfit in the training. Even though our experiment results in AlexNet algorithm becoming overfit, the overall classification performance displayed that the accuracy with the augmentation technique was seen to enhance by 1.08%.

TABLE II. CLASSIFICATION PERFORMANCE OF CNN ALGORITHMS ACROSS TIMBER SPECIES WITH MULTIPLE HYPERPARAMETERS SETTINGS. THE HIGHEST CLASSIFICATION ACCURACY ACROSS TIMBER SPECIES IS INDICATED IN RED

CNN	Hyperparameters		Rubberwood		Merbau		Meranti		KSK	
	Learning rate	Epoch	Ori	AUG	Ori	AUG	Ori	AUG	Ori	AUG
ResNet50	0.001	50	91.85	94.00	89.63	86.74	82.22	92.52	86.67	92.30
		100	89.63	94.59	88.89	90.07	88.89	93.56	86.67	91.41
		200	92.59	94.22	86.67	88.89	91.11	94.07	88.89	92.22
	0.0001	50	82.96	93.33	84.44	88.52	91.85	92.74	84.44	92.67
		100	86.67	92.89	85.19	87.48	92.59	91.85	87.41	93.26
		200	88.15	93.70	83.70	89.19	89.63	92.15	85.93	91.85
GoogLeNet	0.001	50	81.48	92.59	85.19	87.48	91.85	91.19	86.67	92.96
		100	83.70	92.67	85.19	85.63	90.37	91.41	87.41	92.44
		200	85.19	93.04	89.63	89.41	94.07	92.15	85.93	89.70
	0.0001	50	85.19	92.00	75.56	79.56	88.89	91.41	82.96	84.67
		100	88.15	91.56	84.44	81.26	83.70	91.33	87.41	85.48
		200	86.67	92.89	84.44	75.85	82.96	92.52	81.48	85.56
AlexNet	0.001	50	66.67	88.44	81.48	72.07	83.70	84.67	77.78	83.33
		100	68.69	91.56	78.52	69.56	86.67	86.37	84.44	86.07
		200	70.37	89.85	79.26	69.04	85.19	85.63	81.48	11.11
	0.0001	50	82.22	92.81	84.44	83.70	85.93	90.44	86.67	89.56
		100	80.00	91.48	81.48	80.30	88.89	89.11	86.67	90.89
		200	79.26	92.07	80.74	84.30	87.41	89.48	87.41	89.33
ShuffleNet	0.001	50	88.89	93.19	83.70	91.56	87.41	92.52	90.37	86.37
		100	85.19	93.78	84.44	82.15	87.41	92.44	90.37	90.52
		200	88.15	93.56	89.63	87.11	88.89	92.52	88.15	89.33
	0.0001	50	80.74	92.15	79.26	82.52	86.67	90.96	82.96	90.22
		100	79.26	92.74	80.00	81.48	88.15	89.78	83.70	89.85
		200	80.00	90.59	81.48	81.48	92.59	91.33	82.96	90.81
NASNetMobile	0.001	50	84.44	93.56	77.78	88.00	90.37	91.48	83.70	90.89
		100	85.19	94.30	84.44	86.15	92.59	90.81	85.19	92.15
		200	85.19	92.67	80.00	89.33	90.37	92.44	87.41	94.15
	0.0001	50	78.52	93.48	77.78	84.22	92.59	90.15	84.44	89.56
		100	82.22	92.15	79.26	87.85	92.59	89.63	82.96	89.33
		200	76.30	92.30	80.74	89.04	94.07	91.04	88.15	90.15
MobileNetV2	0.001	50	85.19	92.67	85.19	84.37	89.63	88.67	85.93	91.26
		100	82.96	91.85	83.70	82.81	91.11	89.48	85.19	89.26
		200	79.26	92.22	82.22	81.85	88.15	90.37	85.19	89.93
	0.0001	50	82.22	89.41	71.85	80.00	88.89	88.59	84.44	83.85
		100	83.70	92.37	73.33	78.89	89.63	88.74	82.96	84.67
		200	77.78	91.56	74.07	76.89	91.11	89.41	83.70	85.33

In ShuffleNet, applying the augmentation technique in the Rubberwood dataset was seen to improve the accuracy by 13.48% with 0.0001 learning rate and 100 epochs, as displayed in Table II. However, ShuffleNet showed the highest classification performance with regards to the Rubberwood dataset (93.79%), wherein the overall accuracy increased by 4.11% across the timber species by employing data augmentation. Next, the best result of 94.3% was achieved via NASNetMobile by employing 0.001 learning rate and 100 epochs. With regards to the different epoch as well as learning rate combinations, the average classification accuracy can be enhanced by 5.78% by employing the algorithm across the timber species. Most of the defect datasets can achieve accuracy of greater than 94% in NASNetMobile aside from

Merbau dataset that can reach just 89.33%. In line with other data augmentation studies, few augmented datasets could cast an adverse impact on CNN classification performance, like Meranti augmented dataset, which can decrease the performance by 3.03%. With regards to the augmented Rubberwood dataset, MobileNetV2 displayed high classification accuracy enhancement of 12.96% employing 0.001 learning rate and 200 epochs. With the Rubberwood dataset (92.67%), the highest accuracy was displayed, specifying that the algorithm classification performance is enhanced by the augmentation method. After training with 0.0001 learning rate setting, the MobileNetV2 showed decrease in performance in Merbau augmented dataset, similar to the performance of AlexNet. In the Meranti dataset, a majority of performance accuracy degradation was observed from 91.11%

to 89.41% employing 0.0001 learning rate and 200 epochs. However, the overall algorithm performance employing the augmentation method demonstrated enhancement in classification accuracy by 3.63%.

Fig. 1 lists out the overall performance pertaining to multiple CNN algorithms employing both non-augmented and augmented timber defect datasets. These experiments demonstrated that with the help of the augmentation technique, the small dataset issue [34] can be addressed as well as the CNN classification results can be enhanced. Even though all the analysed models demonstrated enhancement in classification performance, the NASNetMobile model gave the best improvement (5.8%). By employing data augmentation, performance enhancement in the range of 1.08–5.78% was demonstrated across the CNN algorithms along with certain fine-tuning with regards to epoch as well as learning rate hyperparameters. Employing the augmentation technique with regards to timber defect identification also increased the average accuracy across the timber species i.e., from 87.78% to 91.84%. The ResNet50 was seen to function well across the

timber species giving an average accuracy of 91.84% along with high performance in terms of defect recognition, when compared with the results pertaining to other CNN algorithms in the timber defect dataset. Fig. 2 on the other hand, displays the validation loss curve of highest accuracy CNN models fine-tuned by two different hyperparameters (learning rate and epoch). Besides, it can be seen from the loss curve that ResNet50 can converge quickly compared to other CNN models. Referring to the t-test in Table III, the CNN classification performance in augmented dataset are significantly better compared to the original dataset with the results demonstrating statistically significant differences between the two datasets except for AlexNet. This evidently displays that the augmentation technique cannot be regarded as a domain specific technique and can be applied for other unexplored timber defect identification domain. Besides, CNN algorithms allow achieving high defect identification performance, which can be leveraged to develop automatic visual inspection system in real-world secondary wood industry processing facilities for optimisation of grading as well as cutting for timber.

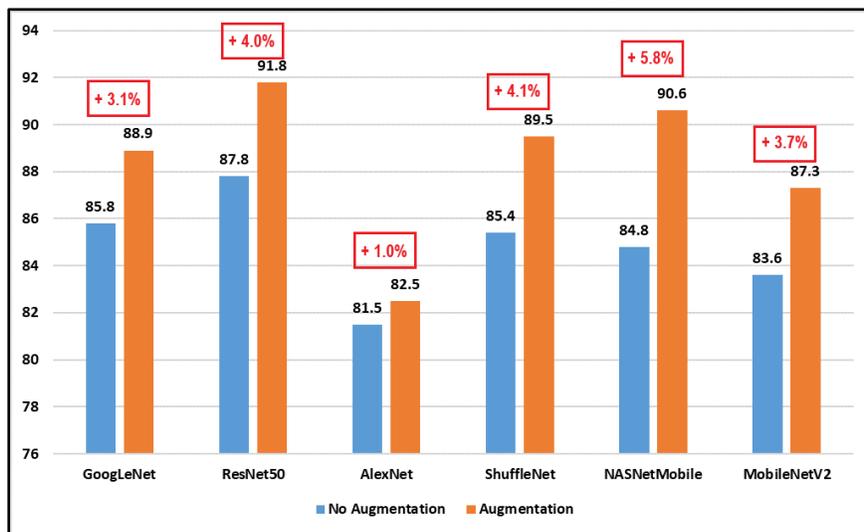


Fig. 1. Overall Performance of CNN Algorithms in both Augmented and Non-Augmented Timber Defect Dataset.

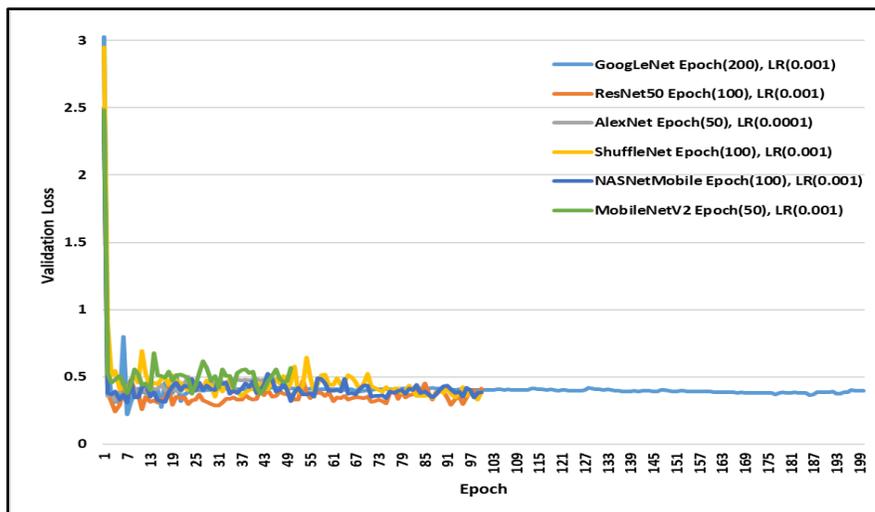


Fig. 2. Loss Curve of CNN Algorithms with Highest Classification Performance.

TABLE III. T-TEST RESULT ON AVERAGE CNN CLASSIFICATION PERFORMANCE IN BOTH AUGMENTED AND NON-AUGMENTED TIMBER DEFECT DATASET

CNN	Original Dataset (\bar{x})	Augmented Dataset (\bar{x})	Sig.	Result
ResNet50	87.78	91.84	.000	Significantly Different
GoogLeNet	85.77	88.95	.014	Significantly Different
AlexNet	81.47	82.55	.767	Significantly Similar
ShuffleNet	85.43	89.54	.001	Significantly Different
NASNetMobile	84.85	90.62	.000	Significantly Different
MobileNetV2	83.64	87.27	.015	Significantly Different

IV. CONCLUSION

This study is aimed at evaluating the effectiveness pertaining to data augmentation technique by using multiple CNN algorithms to identify timber defects across four timber species. The research employs CNN algorithms by implementing transfer learning on ResNet50, GoogLeNet, MobileNetV2, ShuffleNet, NASNetMobile and AlexNet. Evaluation of both data augmentation as well as transfer learning methods was done with various learning rate and epochs to identify the best CNN classification performance for timber species. The result showed data augmentation and transfer learning techniques can be effectively used for searching defect across timber species. The best accuracy could be achieved by employing the ResNet50 model (94.59%) along with optimisation of hyperparameters at learning rate 0.001 and 100 epochs. Our results demonstrate that the augmentation technique can deal with the limited dataset issue as well as enhance the average CNN classification performance by almost 5.78% with regards to the NASNetMobile model. Also, our research study has employed different combination of learning rate as well as epoch, suggesting that a higher number of learning rate and epoch does not necessarily give higher accuracy for CNN model classification. Besides, the research outcomes show that data augmentation as well as CNN algorithms methods with regards to timber defect identification can be used for cutting optimisation as well as industrial timber grading. Also, to further enhance the results, exploring of more complex data augmentation as well as transfer learning framework can be done. The method includes limitations with regards to the requirement of manually adjusting the orientation pertaining to the timber defect images to carry out data augmentation as well as manually label the images pertaining to training CNN algorithms, which may not be regarded as error-free. Future work may include using deep learning for analysis of various kinds of timber defects across different timber species.

ACKNOWLEDGMENT

This research is supported by Universiti Teknikal Malaysia Melaka (UTeM).

REFERENCES

- [1] Y. Yang, X. Zhou, Y. Liu, Z. Hu, and F. Ding, "Wood defect detection based on depth extreme learning machine," *Appl. Sci.*, vol. 10, no. 21, p. 7488, 2020, doi: 10.3390/app10217488.
- [2] V. T. Nguyen, T. Constant, B. Kerautret, I. Debled-Rennesson, and F. Colin, "A machine-learning approach for classifying defects on tree trunks using terrestrial LiDAR," *Comput. Electron. Agric.*, vol. 171, no. February, 2020, doi: 10.1016/j.compag.2020.105332.
- [3] T. H. Chun et al., "Identification of wood defect using pattern recognition technique," *Int. J. Adv. Intell. Informatics*, vol. 7, no. 2, p. 163, Apr. 2021, doi: 10.26555/ijain.v7i2.588.
- [4] S. Soleymani, A. Dabouei, H. Kazemi, J. Dawson, and N. M. Nasrabadi, "Multi-Level Feature Abstraction from Convolutional Neural Networks for Multimodal Biometric Identification," *Proc. - Int. Conf. Pattern Recognit.*, vol. 2018-Augus, no. i, pp. 3469–3476, 2018, doi: 10.1109/ICPR.2018.8545061.
- [5] Y. Zhang, J. Xu, and H. Cheng, "AdaBoost-based conformal prediction with high efficiency," *Int. J. High Perform. Comput. Netw.*, vol. 13, no. 4, pp. 355–365, 2019.
- [6] A. R. de Geus, S. F. d. Silva, A. B. Gontijo, F. O. Silva, M. A. Batista, and J. R. Souza, "An analysis of timber sections and deep learning for wood species classification," *Multimed. Tools Appl.*, vol. 79, no. 45–46, pp. 34513–34529, 2020, doi: 10.1007/s11042-020-09212-x.
- [7] S. Y. Jung, Y. H. Tsai, W. Y. Chiu, J.-S. S. Hu, and C.-T. T. Sun, "Defect detection on randomly textured surfaces by convolutional neural networks," in *IEEE/ASME International Conference on Advanced Intelligent Mechatronics, AIM*, 2018, vol. 2018-July, pp. 1456–1461, doi: 10.1109/AIM.2018.8452361.
- [8] S. Wang, W. Liu, J. Wu, L. Cao, Q. Meng, and P. J. Kennedy, "Training deep neural networks on imbalanced data sets," *Proc. Int. Jt. Conf. Neural Networks*, vol. 2016-October, pp. 4368–4374, 2016, doi: 10.1109/IJCNN.2016.7727770.
- [9] Q. Dong, S. Gong, and X. Zhu, "Imbalanced Deep Learning by Minority Class Incremental Rectification," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 6, pp. 1367–1381, 2018.
- [10] M. Ibrahim, M. Torki, and N. El-Makky, "Imbalanced toxic comments classification using data augmentation and deep learning," in *2018 17th IEEE international conference on machine learning and applications (ICMLA)*, 2018, pp. 875–878.
- [11] J. Hu, W. Song, W. Zhang, Y. Zhao, and A. Yilmaz, "Deep learning for use in lumber classification tasks," *Wood Sci. Technol.*, vol. 53, no. 2, pp. 505–517, 2019, doi: 10.1007/s00226-019-01086-z.
- [12] D. Han, Q. Liu, and W. Fan, "A new image classification method using CNN transfer learning and web data augmentation," *Expert Syst. Appl.*, vol. 95, pp. 43–56, 2018, doi: 10.1016/j.eswa.2017.11.028.
- [13] U. R. Hashim, S. Z. Hashim, and A. K. Muda, "Image collection for non-segmenting approach of timber surface defect detection," *Int. J. Adv. Soft Comput. its Appl.*, vol. 7, no. 1, pp. 15–34, 2015.
- [14] M. Sajjadi, M. Javanmardi, and T. Tasdizen, "Regularization with stochastic transformations and perturbations for deep semi-supervised learning," *Adv. Neural Inf. Process. Syst.*, vol. 29, pp. 1163–1171, 2016.
- [15] Ž. Emeršič, D. Štepec, V. Štruc, and P. Peer, "Training convolutional neural networks with limited training data for ear recognition in the wild," *arXiv Prepr. arXiv1711.09952*, 2017.
- [16] A. Urbonas, V. Raudonis, R. Maskeliūnas, and R. Damaševičius, "Automated identification of wood veneer surface defects using faster region-based convolutional neural network with data augmentation and transfer learning," *Appl. Sci.*, vol. 9, no. 22, p. 4898, 2019, doi: 10.3390/app9224898.
- [17] M. A. Rasyidi, R. Handayani, and F. Aziz, "Identification of batik making method from images using convolutional neural network with limited amount of data," *Bull. Electr. Eng. Informatics*, vol. 10, no. 3, pp. 1300–1307, 2021, doi: 10.11591/eei.v10i3.3035.
- [18] Z. Al-Halah, L. Rybok, and R. Stiefelhagen, "Transfer metric learning for action similarity using high-level semantics," *Pattern Recognit. Lett.*, vol. 72, pp. 82–90, 2016.

- [19] O. Russakovsky et al., "Imagenet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, 2015, doi: 10.1007/s11263-015-0816-y.
- [20] Y. Wang, X. Liu, and C. Yu, "Assisted Diagnosis of Alzheimer's Disease Based on Deep Learning and Multimodal Feature Fusion," *Complexity*, vol. 2021, 2021, doi: 10.1155/2021/6626728.
- [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, doi: 10.1109/CVPR.2018.00716.
- [22] A. Singla, L. Yuan, and T. Ebrahimi, "Food/non-food image classification and food categorization using pre-trained googlenet model," in *Proceedings of the 2nd International Workshop on Multimedia Assisted Dietary Management*, 2016, pp. 3–11.
- [23] I. Z. Mukti and D. Biswas, "Transfer learning based plant diseases detection using ResNet50," in *2019 4th International Conference on Electrical Information and Communication Technology (EICT)*, 2019, pp. 1–6.
- [24] Y. Chu, X. Yue, L. Yu, M. Sergei, and Z. Wang, "Automatic Image Captioning Based on ResNet50 and LSTM with Soft Attention," *Wirel. Commun. Mob. Comput.*, vol. 2020, 2020, doi: 10.1155/2020/8909458.
- [25] A. H. M. Linkon, M. M. Labib, F. H. Bappy, S. Sarker, M. E. Jannat, and M. S. Islam, "Deep Learning Approach Combining Lightweight CNN Architecture with Transfer Learning: An Automatic Approach for the Detection and Recognition of Bangladeshi Banknotes," in *2020 11th International Conference on Electrical and Computer Engineering (ICECE)*, 2020, pp. 214–217, doi: 10.1109/ICECE51571.2020.9393113.
- [26] F. Saxen, P. Werner, S. Handrich, E. Othman, L. Dinges, and A. Al-Hamadi, "Face attribute detection with mobilenetv2 and nasnet-mobile," *Int. Symp. Image Signal Process. Anal. ISPA*, vol. 2019-Septe, no. C, pp. 176–180, 2019, doi: 10.1109/ISPA.2019.8868585.
- [27] C. Buiu, V. R. Dănilă, and C. N. Răduță, "MobileNetV2 ensemble for cervical precancerous lesions classification," *Processes*, vol. 8, no. 5, 2020, doi: 10.3390/PR8050595.
- [28] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4510–4520, doi: 10.1109/CVPR.2018.00474.
- [29] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *European conference on computer vision*, Springer, 2014, pp. 818–833.
- [30] 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.
- [31] S. D. Bimorogo and G. P. Kusuma, "A comparative study of pretrained convolutional neural network model to identify plant diseases on android mobile device," *Int. J. Adv. Trends Comput. Sci. Eng.*, vol. 9, no. 3, pp. 2824–2833, 2020, doi: 10.30534/ijatcse/2020/53932020.
- [32] Y. Ozaki, M. Yano, and M. Onishi, "Effective hyperparameter optimization using Nelder-Mead method in deep learning," *IPSN Trans. Comput. Vis. Appl.*, vol. 9, 2017, doi: 10.1186/s41074-017-0030-7.
- [33] A. Johny and K. N. Madhusoodanan, "Dynamic Learning Rate in Deep CNN Model for Metastasis Detection and Classification of Histopathology Images," *Comput. Math. Methods Med.*, vol. 2021, 2021.
- [34] G.-J. Qi and J. Luo, "Small data challenges in big data era: A survey of recent progress on unsupervised and semi-supervised methods," *IEEE Trans. Pattern Anal. Mach. Intell.*, 2020.

Security Analysis on an Improved Anonymous Authentication Protocol for Wearable Health Monitoring Systems

Gayeong Eom¹

Department of Statistics
Graduate School, Inje University
Gimhae 50834, Republic of Korea

Haewon Byeon^{2*}, Younsung Choi^{3*}

Department of Artificial Intelligence
Inje University, Gimhae 50834
Republic of Korea

Abstract—The wearable health monitoring system (WHMS) plays a significant role in medical experts collecting and using patient medical data. The WHMS is becoming more popular than in the past through mobile devices due to meaningful progress in wireless sensor networks. However, because the data about health used by the WHMS is related to privacy, it has to be protected from malicious access when wirelessly transmitted. Jiang et al. proposed a two-factor suitable for WHMSs using a fuzzy verifier. However, Jiaqing Mo et al. revealed that the protocol proposed by Jiang et al. had various security vulnerabilities and proposed an authentication protocol with improved security and guaranteed anonymity for WHMSs. In this paper, we analyse the authentication protocol proposed by Jiaqing Mo et al. and determine problems with the offline identification, password guessing attacks, operation process bit mismatch, no perfect forward secrecy, no mutual authentication and insider attacks.

Keywords—Authentication protocol; health status; physiological data; security analysis; WHMS

I. INTRODUCTION

Electronic health system keeps Wireless communication, authentication protocol, sensors using low-power, and security solution on authentication protocol [1-8] safe. Wireless sensor networks (WSNs), which play a significant role in e-health, detect, measure, collect or record patient information on a medical server for physician diagnosis. The wearable health monitoring system (WHMS) has received considerable attention regarding its movability, adaptability and operation cost [9, 10, 11, 12]. The WHMS detects, measures, and collects patient physiological data with the WSN inserted or embed within the patient's body. In addition, after monitoring the health status, information is transmitted through wireless channels to medical-related institutions to help manage it. Remote doctors can evaluate the health status through such data as the heart rate, blood pressure and body temperature.

The WHMS is simple and efficient for medical professionals, and patients receive many benefits from the WHMS. However, the detected data are transmitted over an unsafe wireless channel; thus, concerns exist regarding security and privacy problems. Therefore, a robust certified mechanism must be designed to protect the physiological data for patients whose security is critical. If an adversary modifies the data for

a patient, the doctor will misdiagnose the patient based on incorrect data. In addition, revealed data are highly likely to be used by malicious and illegal purposes. Medical personnel must authenticate that they are normal users before accessing patient's physiological data from the wearable sensor of the patient to prevent this. Even if the adversary eavesdropped on the message through the gateway of the WHMS, their identity and passwords must not be disclosed. A session key must be calculated between the sensor node and the medical personnel on the patient's body for future secure communication.

Kumar et al. [13] studied a user authentication protocol in 2012 to monitor patient physiological data in the medical WSN E-SAP and argued that the protocol was safe on the known attacks. But He et al. [14] and Khan and Kumari [15] found security vulnerabilities, such as lack of user anonymity and password guessing attack in the plans by Kumar et al. and presented improved versions. Li et al. and Wu et al., Mir et al. [16-18] individually found out that the plan by He et al. [14] has security problems, such as offline guessing attacks, denial-of-service (DoS) attacks, most attacks, and sensor node capture attacks. They proposed an enhanced version that is safer than the previous proposal to compensate for these loopholes. Das et al. [20] pointed out various security flaws such as lack of user anonymity, privileged insider attacks and sensor capture attacks in the protocol by Li et al. [21] and proposed an improved framework based on biometric recognition. Amin et al. [19] proposed a mutually authentication protocol providing user's anonymity in the WHMS and stated that the system was secure against already known various attacks. But Jiang et al. [22] revealed that the protocol has various vulnerabilities like as unsynchronised attacks, sensor key exposure and stolen mobile device attacks. Jiang et al. proposed an enhanced authentication protocol using smart card and password [22, 23]. Their protocol used square surplus, fuzzy validator [24] and timestamp mechanisms to ensure the plan by Amin et al. In addition, as a result of a security analysis, their plan achieved the desired security function.

Separately, Challa et al. [25] claimed an enhanced 3-factor (cryptography, smart card and biometric) authentication scheme for healthcare WSNs to enhance the scheme's security proposed by Liu and Chung [26]. However, this method, which requires the user to communicate directly with the remote sensor, greatly increases the sensor power consumption and

*Corresponding Author.

rapidly reduces its lifespan. Therefore, their systems cannot be applied to healthcare WSNs. Ali et al. [27] proposed a 3-factor protocol providing anonymity in the plan by Amin et al. [19] to frustrate security threats, such as user impersonation attacks, offline password guessing attacks and known session key temporary information attacks. Shen et al. [28] presented a multilayer authentication protocol using ECC in WBANs (wireless body area networks) to improve authentication's security and compute group key generation between sensors and mobile devices. Li et al. [29] proposed an efficient authentication scheme for a centralised WBAN organized two hops while maintaining anonymity and nonconnectedness in data transmission. And Shen et al. [30] proposed an ECC-based authentication protocol using public key signature scheme for WBAN. But according to [31, 32], their authentication scheme type with only two round messages is likely to fail in perfect forward secrecy.

Jiaqing Mo et al. analysed the protocol proposed by Jiang et al. [22] and discovered that Jiang et al.' protocol was not safe as their proven. Jiang et al.' scheme provides fuzzy verifiers to block offline password guessing attacks, their systems were still vulnerable to authoritative insider attacks, leading to user impersonation attacks. Unfortunately, the plan by Jiang et al. [22] is subject to KSSTI attacks; thus, their protocols are as vulnerable to sensor key disclosure as before. In addition, their protocols struggle with DoS attacks. In addition, Jiaqing Mo et al. implement an authentication scheme with improved security and guaranteed anonymity for WHMSs to solve this problem. However, in this paper, we analyse the authentication protocol proposed by Jiaqing Mo et al. and discovered problems with the offline identification, password guessing attacks, operation process bit mismatch, no perfect forward secrecy, no mutual authentication and insider attacks.

The rest of this paper is organised as follows. Section 2 describes the terms and adversary models used in this paper. Section 3 analyses the operation process of an authentication protocol with improved security and guaranteed anonymity for the WHMS proposed by Jiaqing Mo et al. Section 4 describes the vulnerabilities found by conducting a stability analysis on the protocol proposed by Jiaqing Mo et al. Finally, in Section 5, we conclude this paper.

II. RELATED RESEARCH

A. Summary of Symbol

Symbols used in the paper's operation process are shown in Table I.

B. Adversary Model

An adversary's capabilities are essential part of an adversary model. In this paper, it is assumed that the adversary has the following capabilities.

- An adversary can completely control open channels like as inserting, intercepting, eavesdropping, deleting, and modifying exchanged messages through open channels [33, 34].
- An adversary can find out all data (i.e. secret key and random number) stored in the mobile device when adversary acquire user's lost mobile device [35, 36].

- An adversary can estimate the ID_i and PW_i offline by listing pairs in Cartesian product $D_{ID} \times D_{PW}$ within polynomial time. Here, the D_{ID} represents identity space and D_{PW} is password space [31, 37].
- The secret key and random numbers party are suitably large so they overcome adversary from successfully guessing accurate data within polynomial time.
- The inside adversary may get a user's registration request message, and the insider may access the verification table [38, 39].

TABLE I. SUMMARY OF SYMBOL

Symbol	Meaning
U_i	Medical professional
ID_i	U_i 's identity
PW_i	U_i 's password
S_j	The j th sensor node
SID_j	S_j 's identity
GWN	Gateway
K	GWN's secret key
MD	The mobile device
R_1, R_2 and R_3	Random nonce produced by U_i , GWN, and S_j , respectively
\oplus	Bitwise XOR operation
\parallel	Concatenation
$h()$	One-way hash function

III. OPERATION PROCESS OF THE PROTOCOL PROPOSED BY JIAQING MO ET AL

Jiaqing Mo et al.'s proposed protocol consists of five stages: setting, medical expert registration, patient registration, login and authentication, and password change.

A. Setting Step

The registration center GWN selects two large primes p and q , computes $n = pq$, and maintains the private key (p, q) .

B. Medical Professional Registration Step

1) U_i inserts own ID_i and PW_i , a random-nonce r_i , and computes $HPW_i = h(r_i \oplus PW_i)$; then send $\{ID_i, HPW_i\}$ to gateway through a secure channel.

2) After receiving user's registration request, GWN selects $m \in [2^4, 2^8]$, a random-nonce R_i , computes $Reg_i = h(h(ID_i \parallel R_i \parallel HPW_i) \bmod m)$, $A_i = R_i \oplus HPW_i$, $B_i = h(ID_i \parallel R_i \parallel K)$, and $C_i = B_i \oplus h(ID_i \parallel R_i \parallel HPW_i)$. Reg_i is a fuzzy verifier. Thereafter, GWN transmits $\{Reg_i, A_i, C_i, m, n, h()\}$ to U_i using a secure channel.

3) When U_i receive GWN' message, U_i computes $A_i^* = A_i \oplus h(ID_i \parallel r_i)$, $D_i = r_i \oplus h(h(ID_i \parallel PW_i) \bmod m)$ and updates MD to $\{Reg_i, A_i^*, C_i, D_i, m, n, h()\}$.

C. Patient Registration Step

This step is almost identical to Jiang's plan [22].

- 1) The patient delivers the ID to the registration center.
- 2) Select the appropriate sensor kit from the registration center and assigns a professional.
- 3) The registration center calculates $SK_{GW-sj} = h(SID_j \parallel K)$ for S_j as a secret key between GWN and sensor node. And the registration center delivers the patient's significant information to the designated specialist.

D. The Login and Authentication Step

Through this step, this protocol will be able to provide mutual authentication and generate session keys between U_i and S_j for future communication.

1) U_i chooses own ID_i and PW_i , and MD calculates $r_i = D_i \oplus h(h(ID_i \parallel PW_i) \text{ mod } m)$, $HPW_i = h(r_i \oplus PW_i)$, $A_i = A_i^* \oplus h(ID_i \parallel r_i)$, $R_i^* = A_i \oplus HPW_i$, $Reg_i^* = h(h(ID_i \parallel R_i^* \parallel HPW_i^*) \text{ mod } m)$, and tests $Reg_i^* = Reg_i$. If false, MD selects a random number R_1 and calculates $B_i^* = C_i \oplus h(ID_i \parallel R_i \parallel HPW_i)$, $CID_i = (ID_i \parallel R_1 \parallel R_i^* \parallel SID_j)^2 \text{ mod } n$, $M_1 = h(ID_i \parallel B_i^* \parallel R_1 \parallel T_1)$, then transfers $msg_1 = \{CID_i, M_1, T_1\}$ to GWN. T_1 is the current timestamp.

2) After receiving the login request msg_1 , the GWN decrypts CID_i with (p, q) to obtain $(ID_i^*, R_i^*, R_1^*, T_1)$ and confirms the freshness of the timestamp T_1 . If the confirmation fails, GWN stops the session. Otherwise, GWN calculates $B_i' = h(ID_i \parallel R_i \parallel K)$ and $M_1^* = h(ID_i \parallel B_i' \parallel R_1 \parallel T_1)$ and then tests $M_1^* = M_1$. If inequality persists, GWN stops the procedure. Otherwise, GWN computes $SK_{GW-sj} = h(SID_j \parallel K)$, selects a random nonce R_2 , and computes $M_2 = h(ID_i^* \parallel R_1^* \parallel R_i)$, $M_3 = h(h(M_2 \parallel "1") \parallel SK_{GW-sj} \parallel R_2 \parallel T_2)$, $M_4 = M_2 \oplus h(SK_{GW-sj} \parallel T_2)$, and $M_5 = R_2 \oplus h(SK_{GW-sj} \parallel SID_j \parallel T_2)$. Finally, GWN sends $msg_2 = \{M_3, M_4, M_5, T_2\}$ to S_j .

3) When receiving msg_2 from GWN, firstly S_j checks the validity of T_2 . If it is not fresh, S_j stops next procedure. If it is fresh, S_j calculates $R_2' = M_5 \oplus (SK_{GW-sj} \parallel SID_j \parallel T_2)$ and $M_2 = M_4 \oplus h(SK_{GW-sj} \parallel T_2)$ and tests $M_3 = h(h(M_2 \parallel "1") \parallel SK_{GW-sj} \parallel R_2' \parallel T_2)$. If it is false, S_j terminates the session. Otherwise, S_j selects a random number R_3 and calculates $SK = h(M_2' \parallel R_2' \parallel R_3)$, $M_6 = h(SK \parallel R_3 \parallel SK_{GW-sj})$, and $M_7 = h(R_2' \parallel T_3) \oplus R_3$, where T_3 is the current timestamp. Then, S_j transfers $msg_3 = \{M_6, M_7, T_3\}$ to GWN.

4) When msg_3 is a received from S_j , the GWN confirms the validity of T_3 firstly. If timestamp is fresh, GWN terminates next procedure. Otherwise, GWN calculates $R_3' = M_7 \oplus h(R_2' \parallel T_3)$, $SK' = h(M_2 \parallel R_2 \parallel R_3')$, and $M_6' = h(SK' \parallel R_3' \parallel SK_{GW-sj})$ and examines whether $M_6' = M_6$ holds. If they are same, GWN computes $M_8 = R_2 \oplus h(ID_i^* \parallel R_1^*)$, $M_9 = R_3 \oplus h(ID_i^* \parallel R_2^*)$, and $M_{10} = h(ID_i^* \parallel SK' \parallel R_3 \parallel T_4)$ and transfers $msg_4 = \{M_8, M_9, M_{10}, T_4\}$ to U_i . Here, T_4 is the current timestamp.

5) U_i receives msg_4 from GWN and examines the timestamp T_4 . If timestamp is not fresh, U_i stops next procedure. Otherwise, U_i calculates $R_2' = M_8 \oplus h(ID_i \parallel R_1)$, $R_3' = M_9 \oplus h(ID_i \parallel R_2')$, and $SK^* = h(h(ID_i \parallel R_1 \parallel R_i') \parallel$

$R_2' \parallel R_3')$ and checks whether $M_{10} = h(ID_i \parallel SK^* \parallel R_3' \parallel T_4)$ holds. If they are not same, U_i terminates the current connection. If they are same, U_i can believe that both GWN and S_j are believable. Then U_i and S_j can proceed with secure communication in the future by using the session key. The login and authentication steps are summarized in Fig. 1.

IV. SECURITY ANALYSIS OF JIAQING MO ET AL'S PROTOCOL

This paper analyzed the operation process of Jiang et al.'s protocol and found various vulnerability as off-line ID, PW guessing attack, operation process bit mismatch, no perfect forward secrecy, no mutual authentication and insider attack.

A. Off-line ID, PW Guessing Attack

According to Jiaqing Mo et al.'s proposed protocol, when an adversary acquires a MD, the adversary can extract information stored in the MD and then find out the user's ID and PW. The information of $\{Reg_i, A_i, C_i, m, n, h()\}$ is sent to the MD through the GWN security channel. Thereafter, the MD calculates and updates $A_i^* = A_i \oplus h(ID_i \parallel r_i)$ and $D_i = r_i \oplus h(h(ID_i \parallel PW_i) \text{ mod } m)$. Finally, information of $\{Reg_i, A_i^*, C_i, D_i, m, n, h()\}$ is stored in the MD. Assuming that an adversary found out this through a physical analysis method, an ID and password can be derived through the formula of $Reg_i^* = h(h(ID_i \parallel R_i^* \parallel HPW_i^*) \text{ mod } m)$.

$$\begin{aligned}
 Reg_i^* &= h(h(ID_i \parallel R_i^* \parallel HPW_i^*) \text{ mod } m) \\
 &= h(h(ID_i \parallel A_i \oplus HPW_i^* \parallel h(r_i \oplus PW_i)) \text{ mod } m) \\
 &= h(h(ID_i \parallel A_i^* \oplus h(ID_i \parallel r_i) \oplus HPW_i^* \\
 &\quad \parallel h(r_i \oplus PW_i)) \text{ mod } m) \\
 &= h(h(ID_i \parallel A_i^* \oplus h(ID_i \parallel D_i \oplus h(h(ID_i \\
 &\quad \parallel PW_i) \text{ mod } m)) \oplus h(r_i \oplus PW_i) \\
 &\quad \parallel h(r_i \oplus PW_i)) \text{ mod } m) \\
 &= h(h(ID_i \parallel A_i^* \oplus h(ID_i \parallel D_i \oplus h(h(ID_i \\
 &\quad \parallel PW_i) \text{ mod } m)) \oplus h(D_i \oplus h(h(ID_i \\
 &\quad \parallel PW_i) \text{ mod } m) \oplus PW_i) \parallel h(D_i \oplus h(h(ID_i \\
 &\quad \parallel PW_i) \text{ mod } m) \oplus PW_i)) \text{ mod } m)
 \end{aligned}$$

Summarizing the above formula, the adversary will be aware of the information $\{A_i^*, D_i, m, h()\}$ except for the ID and PW. The adversary repeatedly performs verification while continuing to change until the user's ID and PW are found. Ultimately, the user's exact ID and PW can be found. The process of ID, PW guessing attack is summarized in Fig. 2.

B. Operation Process Bit Mismatch

In Jiaqing Mo et al.'s protocol, XOR operations are widely used, and XOR operations must have the same number of bits. However, in Jiaqing Mo et al.'s protocol, there may be a problem with the XOR operation because the number of bits does not match during the XOR operation. A hash function is a function that receives a message having an arbitrary length and outputs a hash value of a fixed length. Keys are used for cryptographic algorithms, but hash functions do not use keys, so the same output is always produced for the same input. The purpose of using these hash functions is to provide integrity to detect errors or alterations in messages.

$$HPW_i = h(r_i \oplus PW_i)$$

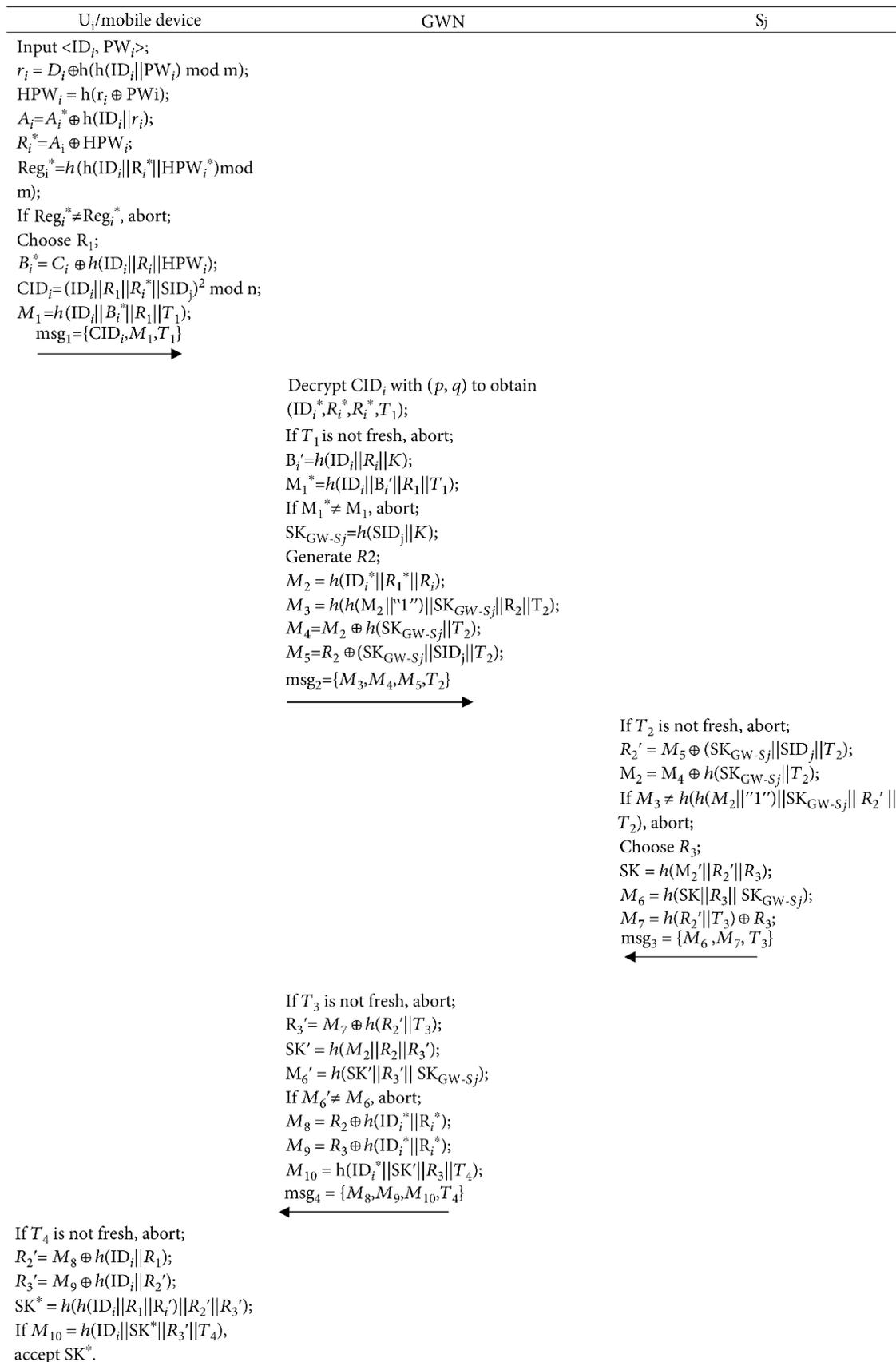


Fig. 1. The Login and Authentication Phase of Jiaqing Mo Et Al.'s Protocol.

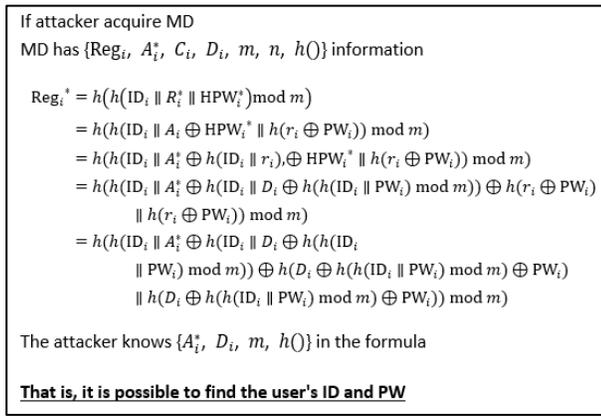


Fig. 2. Process of ID, PW Guessing Attack.

Random nonce values used in the formula usually use large random numbers of 128 bits or more, but the length of the password is very short compared to Random nonce. That is, the length of the random nonce and the length of the password cannot be the same. Therefore, there may be a problem with the XOR operation due to inconsistency in the number of bits in Jiaqing Mo et al.'s protocol.

C. No Perfect Forward Secrecy

The fact that the Perfect Forward Secrecy is met means that even if one of the important master keys in the protocol is exposed, the previous session key cannot be determined. However, in this protocol, the exposure of the (p, q) value, one of the unchanged long-term keys, does not meet the Perfect Forward Secrecy because it can identify not only future session keys but also previously used session keys. That is, assuming that the adversary has found out (p, q), it is possible to calculate the previous session key used between the mobile device and S_j.

1) The adversary has exposed (p, q) values and previous communication contents (CID_i of msg₁, M₈ and M₉ of msg₄) between the user and GWN and S_j. The adversary may decrypt the CID_i of the login request msg₁ as (p, q), and the adversary may find out ID_i^{*}, R_i^{*}, R₁^{*} and T₁.

2) In addition, the adversary may calculate R₂' using M₈, ID_i and R₁.

$$R_2' = M_8 \oplus h(\text{ID}_i \parallel R_1).$$

3) In addition, R₃' may be calculated using M₉, ID_i and R₂'.

$$R_3' = M_9 \oplus h(\text{ID}_i \parallel R_2').$$

4) Finally, the adversary may calculate the session key SK* by using the ID_i, R_i, R₁, R₂' and R₃' obtained so far.

$$\text{SK}^* = h(h(\text{ID}_i \parallel R_1 \parallel R_i') \parallel R_2' \parallel R_3').$$

Since long-term key (p, q) is a key that does not change after it is generated, it is a serious problem that the previous session key is exposed because it does not satisfy the Perfect Forward Secrecy when (p, q) is exposed.

D. No Mutual Authentication

Mutual authentication means that all components of the authentication protocol authenticate with each other. In the present protocol, U_i, GWN, S_j authenticates using M₁, M₃, M₆, M₁₀. Through four messages, mutual authentication between U_i and GWN and mutual authentication between GWN and S_j are provided, but there is a problem of not providing mutual authentication between U_i and S_j. The mutual authentication process is as follows.

1) GWN verifies the authentication of U_i using ID_i and M₁ = h(ID_i || B_i^{*} || R₁ || T₁) having the secret key K of GWN. When M₁ that U_i has is transmitted to GWN, GWN calculates B_i^{*} = h(ID_i || R_i || K) and M₁^{*} = h(ID_i || B_i^{*} || R₁ || T₁). When M₁ and M₁^{*} match, GWN authenticates that U_i is a normal user.

2) When the consistency is confirmed, S_j confirms the authentication of GWN using M₂ = h(ID_i^{*} || R₁^{*} || R_i) and SK_{GW-Sj} = h(SID_j || K). The authentication is confirmed by comparing M₃ = h(h(M₂ || "1") || SK_{GW-Sj} || R₂ || T₂) and h(h(M₂ || "1") || SK_{GW-Sj} || R₂' || T₂) having session key SK_{GW-Sj} of GWN and S_j.

3) When authentication is confirmed, GWN checks the consistency between M₆ = h(SK || R₃ || SK_{GW-Sj}) and M₆' = h(SK' || R₃' || SK_{GW-Sj}) to confirm the authentication of S_j.

4) Finally, if M₁₀ = h(ID_i^{*} || SK' || R₃ || T₄) matches h(ID_i || SK* || R₃' || T₄), U_i authenticates GWN.

GWN authenticates U_i through M₁, and S_j authenticates GWN through M₃. Through M₆, GWN authenticates S_j, and U_i authenticates GWN through M₁₀. That is, U_i and GWN, GWN and S_j are mutually authenticated, but in this protocol, mutual authentication between U_i and S_j is not provided. In order to create an authentication protocol with improved security, the authentication protocol will be safer only when U_i and S_j are also mutually authenticated. Fig. 3 describes in detail how the mutual authentication process is performed.

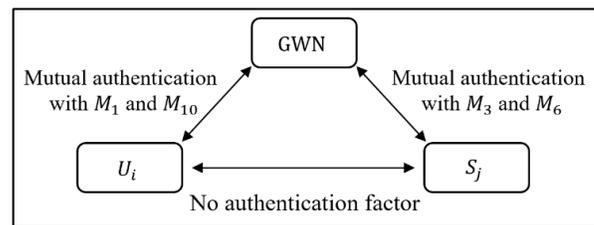


Fig. 3. Mutual Authentication Process.

E. Insider Attack

Even an insider of GWN should not be able to pretend to be a normal user by utilizing the information obtained in the process of verifying the user's authentication information in the MD authentication step. However, in the protocol proposed by Jiaqing Mo et al., there is a problem that insiders can disguise themselves as normal users using only {ID_i^{*}, R_i^{*}}. In this protocol, in the process of calculating the user's authentication information, an internal adversary can find out the user's {ID_i^{*}, R_i^{*}} information that authenticates with the GWN's secret key K. Based on this information, an internal adversary can

REFERENCES

succeed in authentication under the guise of a normal user, and a session key can also be calculated. Fig. 4 shows the protocol authentication process and the adversary calculating the session key.

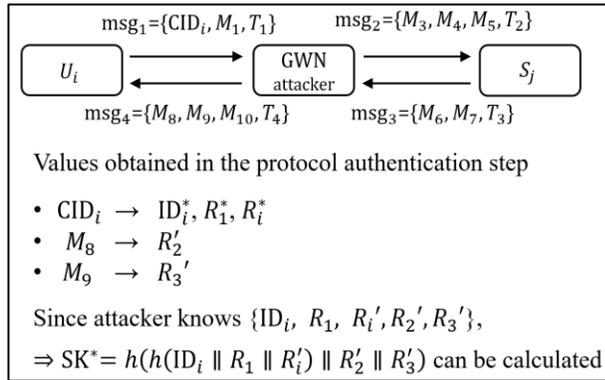


Fig. 4. The Adversary's Session Key Calculation Process.

Among $msg_1 = \{CID_i, M_1, T_1\}$ transmitted to GWN by an insider, $CID_i = (ID_i || R_1 || R_i^* || SID_j)^2 \bmod n$ may calculate using the unchanged values ID_i^* and R_i^* obtained by an insider adversary. In the case of M_1, B_i^* of $M_1 = h(ID_i || B_i^* || R_1 || T_1)$ may be found using information of $B_i' = h(ID_i || R_i || K)$ known by an internal adversary. Since the T_1 value can also be generated by the internal adversary at the current time, M_1 can be calculated. This allows an internal adversary to succeed in authentication under the guise of a user with only the information received from GWN. An insider adversary who succeeds in logging in receives $\{M_8, M_9, M_{10}, T_4\}$ information through $msg_1, msg_2, msg_3, msg_4$. The insider adversary must calculate information of R_2' and R_3' to compute the session key. Since the insider adversary has all the information in $R_2' = M_8 \oplus h(ID_i || R_1), R_2'$ may be calculated, and $R_3' = M_9 \oplus h(ID_i || R_2')$ may be calculated using R_2' . An insider adversary may calculate $SK^* = h(h(ID_i || R_1 || R_i') || R_2' || R_3')$ because it has all the information of $\{ID_i, R_1, R_i', R_2', R_3'\}$ necessary for calculating the session key. As a result, authentication can be successful under the guise of a normal user only with the information possessed by the insider adversary.

V. CONCLUSION

In this paper, a security analysis was conducted after explaining the operation process of an authentication protocol with improved security and guaranteed anonymity for the WHMS proposed by Jiaqing Mo et al. The protocols proposed by Jiaqing Mo et al. have vulnerabilities in offline identification, password guessing attacks, operation process bit mismatch, no perfect forward secrecy, no mutual authentication and insider attack problems.

ACKNOWLEDGMENT

This research was funded by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, grant number "2018R1D1A1B07041091, 2021S1A5A8062526", and "2022 Development of Open-Lab based on 4P in the Southeast Zone".

- [1] R. Amin and G. P. Biswas, A secure three-factor user authentication and key agreement protocol for TMIS with user anonymity, *J. Med. Syst.*, vol. 39, no. 8, pp. 1-19, 2015. <https://doi.org/10.1007/s10916-015-0258-7>.
- [2] S. A. Chaudhry, H. Naqvi, and M. K. Khan, An enhanced lightweight anonymous biometric based authentication scheme for TMIS, *Multimed. Tools Appl.* vol. 77, no. 5, pp. 5503-5524, 2018. <https://doi.org/10.1007/s11042-017-4464-9>.
- [3] F. Wei, P. Vijayakumar, J. Shen, R. Zhang, and L. Li, A provably secure password-based anonymous authentication scheme for wireless body area networks, *Comput. Electr. Eng.* vol. 65, pp. 322-331, 2018. <https://doi.org/10.1016/j.compeleceng.2017.04.017>.
- [4] X. Liu, C. Jin, and F. Li, An improved two-layer authentication scheme for wireless body area networks, *J. Med. Syst.* vol. 42, no. 8, pp. 1-14, 2018. <https://doi.org/10.1007/s10916-018-0990-x>.
- [5] L. Zhang, Y. Zhang, S. Tang, and H. Luo, Privacy protection for e-health systems by means of dynamic authentication and three-factor key agreement, *IEEE Trans. Ind. Electron.* vol. 65, no. 3, pp. 2795-2805, 2018. doi: 10.1109/TIE.2017.2739683.
- [6] O. Mir and M. Nikooghadam, A secure biometrics based authentication with key agreement scheme in telemedicine networks for e-health services, *Wirel. Pers. Commun.* vol. 83, no. 4, pp. 2439-2461, 2015. <https://doi.org/10.1007/s11277-015-2538-4>.
- [7] Q. Jiang, M. K. Khan, X. Lu, J. Ma, and D. He, A privacy preserving three-factor authentication protocol for e-health clouds, *J. Supercomput.* vol. 72, no. 10, pp. 3826-3849, 2016. <https://doi.org/10.1007/s11227-015-1610-x>.
- [8] Y. K. Ever, Secure-anonymous user authentication scheme for e-healthcare application using wireless medical sensor networks, *IEEE Syst J*, vol. 13, no. 1, pp. 456-467, 2019. doi: 10.1109/JSYST.2018.2866067.
- [9] M. M. Baig, H. Gholamhosseini, and M. J. Connolly, A comprehensive survey of wearable and wireless ECG monitoring systems for older adults, *Med. Biol. Eng. Comput.* vol. 52, no. 5, pp. 485-495, 2013. <https://doi.org/10.1007/s11517-012-1021-6>.
- [10] Z. Yang, Q. Zhou, L. Lei, K. Zheng, and W. Xiang, An IoTcloud based wearable ECG monitoring system for smart healthcare, *J. Med. Syst.* vol. 40, no. 12, pp. 1-11, 2016. <https://doi.org/10.1007/s10916-016-0644-9>.
- [11] Y. Yin, H. Jiang, S. Feng et al., Bowel sound recognition using SVM classification in a wearable health monitoring system, *Sci. China Inf. Sci.* vol. 61, no. 8, pp. 1-3, 2018. <https://doi.org/10.1007/s11432-018-9395-5>.
- [12] V. Trovato, C. Colleoni, A. Castellano, and M. R. Plutino, The key role of 3-glycidoxypropyltrimethoxysilane sol-gel precursor in the development of wearable sensors for health monitoring, *J. Sol-Gel Sci. Technol.* vol. 87, no. 1, pp. 27-40, 2018. <https://doi.org/10.1007/s10971-018-4695-x>.
- [13] P. Kumar, S. G. Lee, and H. J. Lee, E-sap: efficient-strong authentication protocol for healthcare applications using wireless medical sensor networks, *Sensors*, vol. 12, no. 2, pp. 1625-1647, 2012. <https://doi.org/10.3390/s120201625>.
- [14] D. He, N. Kumar, J. Chen, C.-C. Lee, N. Chilamkurti, and S.-S. Yeo, Robust anonymous authentication protocol for health-care applications using wireless medical sensor networks, *Multimed. Syst.* vol. 21, no. 1, pp. 49-60, 2015. <https://doi.org/10.1007/s00530-013-0346-9>.
- [15] M. K. Khan and S. Kumari, An improved user authentication protocol for healthcare services via wireless medical sensor networks, *Int. J. Distrib. Sens. Netw.* vol. 10, no. 4, pp. 347169, 2014. <https://doi.org/10.1155/2014/347169>.
- [16] F. Wu, L. Xu, S. Kumari, and X. Li, An improved and anonymous two-factor authentication protocol for health-care applications with wireless medical sensor networks, *Multimed. Syst.* vol. 23, no. 2, pp. 195-205, 2015. <https://doi.org/10.1007/s00530-015-0476-3>.
- [17] O. Mir, J. Munilla, and S. Kumari, Efficient anonymous authentication with key agreement protocol for wireless medical sensor networks, *Peer Peer Netw Appl.* vol. 10, no. 1, pp. 79-91, 2015. <https://doi.org/10.1007/s12083-015-0408-1>.

- [18] C. T. Li, C. C. Lee, and C. Y. Weng, A secure cloud-assisted wireless body area network in mobile emergency medical care system, *J. Med. Syst.* vol. 40, no. 5, pp. 1-15, 2016. <https://doi.org/10.1007/s10916-016-0474-9>.
- [19] R. Amin, S. H. Islam, G. P. Biswas, M. K. Khan, and N. Kumar, A robust and anonymous patient monitoring system using wireless medical sensor networks, *Future Gener. Comput. Syst.* vol. 80, pp. 483-495, 2016. <https://doi.org/10.1016/j.future.2016.05.032>.
- [20] A. K. Das, A. K. Sutrala, V. Odelu, and A. Goswami, A secure smartcard-based anonymous user authentication scheme for healthcare applications using wireless medical sensor networks, *Wirel. Pers. Commun.* vol. 94, no. 3, pp. 1899-1933, 2017. <https://doi.org/10.1007/s11277-016-3718-6>.
- [21] X. Li, J. Niu, S. Kumari, J. Liao, W. Liang, and M. K. Khan, A new authentication protocol for healthcare applications using wireless medical sensor networks with user anonymity, *Secur. Commun. Netw.* vol. 9, no. 15, pp. 2643-2655, 2016. <https://doi.org/10.1002/sec.1214>.
- [22] Q. Jiang, J. Ma, C. Yang, X. Ma, J. Shen, and S. A. Chaudhry, Efficient end-to-end authentication protocol for wearable health monitoring systems, *Comput. Electr. Eng.* vol. 63, pp. 182-195, 2017. <https://doi.org/10.1016/j.compeleceng.2017.03.016>.
- [23] K. H. Rosen, *Elementary Number Theory and Its Applications*, Addison-Wesley, Reading, MA, USA, 1988.
- [24] C.-G. Ma, D. Wang, and S. D. Zhao, Security flaws in two improved remote user authentication schemes using smart cards, *Int. J. Commun. Syst.* vol. 27, no. 10, pp. 2215-2227, 2015. <https://doi.org/10.1002/dac.2468>.
- [25] S. Challa, A. K. Das, V. Odelu et al., An efficient ECC-based provably secure three-factor user authentication and key agreement protocol for wireless healthcare sensor networks, *Comput. Electr. Eng.* vol. 69, pp. 534-554, 2018. <https://doi.org/10.1016/j.compeleceng.2017.08.003>.
- [26] C. H. Liu and Y. F. Chung, Secure user authentication scheme for wireless healthcare sensor networks, *Comput. Electr. Eng.* vol. 59, pp. 250-261, 2017. <https://doi.org/10.1016/j.compeleceng.2016.01.002>.
- [27] R. Ali, A. K. Pal, S. Kumari, A. K. Sangaiah, X. Li, and F. Wu, An enhanced three factor based authentication protocol using wireless medical sensor networks for healthcare monitoring, *J. Ambient Intell. Humaniz. Comput.* pp. 1-22, 2018. <https://doi.org/10.1007/s12652-018-1015-9>.
- [28] J. Shen, S. Chang, J. Shen, Q. Liu, and X. Sun, A lightweight multi-layer authentication protocol for wireless body area networks, *Future Gener. Comput. Syst.* vol. 78, pp. 956-963, 2016. <https://doi.org/10.1016/j.future.2016.11.033>.
- [29] X. Li, M. H. Ibrahim, S. Kumari, A. K. Sangaiah, V. Gupta, and K. K. R. Choo, Anonymous mutual authentication and key agreement scheme for wearable sensors in wireless body area networks, *Comput. Netw. Syst.* vol. 129, pp. 429-443, 2017. <https://doi.org/10.1016/j.comnet.2017.03.013>.
- [30] J. Shen, Z. Gui, S. Ji, J. Shen, H. Tan, and Y. Tang, Cloud-aided lightweight certificateless authentication protocol with anonymity for wireless body area networks, *J. Netw. Comput. Appl.* vol. 106, pp. 117-123, 2018. <https://doi.org/10.1016/j.jnca.2018.01.003>.
- [31] D. Wang and P. Wang, Two birds with one stone: two-factor authentication with security beyond conventional bound, *IEEE Trans. Dependable Secure Comput.* vol. 15, no. 4, pp. 708-722, 2016. doi: 10.1109/TDSC.2016.2605087.
- [32] H. Krawczyk, HMQV: a high-performance secure Diffie-Hellman protocol, in *Advances in Cryptology – CRYPTO 2005*, V. Shoup, Ed., vol. 3621 of *Lecture Notes in Computer Science*, Berlin, Heidelberg, pp. 546-566, 2005. https://doi.org/10.1007/11535218_33.
- [33] C. Wang, G. Xu, and J. Sun, An enhanced three-factor user authentication scheme using elliptic curve cryptosystem for wireless sensor networks, *Sensors*, vol. 17, no. 12, pp. 2946, 2017. <https://doi.org/10.3390/s17122946>.
- [34] R. Amin, S. H. Islam, G. P. Biswas, M. K. Khan, L. Leng, and N. Kumar, Design of an anonymity-preserving three-factor authenticated key exchange protocol for wireless sensor networks, *Comput. Netw.* vol. 101, pp. 42-62, 2016. <https://doi.org/10.1016/j.comnet.2016.01.006>.
- [35] S. Mangard, E. Oswald, and T. Popp, *Power Analysis Attacks: Revealing the Secrets of Smart Cards*, Springer Science & Business Media, 2010.
- [36] T. H. Kim, C. K. Kim, and I. H. Park, Side channel analysis attacks using AM demodulation on commercial smart cards with SEED, *J. Syst. Softw.* vol. 85, no. 12, pp. 2899-2908, 2012. <https://doi.org/10.1016/j.jss.2012.06.063>.
- [37] Q. Jiang, S. Zeadally, J. Ma, and D. He, Lightweight threefactor authentication and key agreement protocol for internet-integrated wireless sensor networks, *IEEE Access*, vol. 5, pp. 3376-3392, 2017. doi: 10.1109/ACCESS.2017.2673239.
- [38] Y. Choi, Y. Lee, J. Moon, and D. Won, Security enhanced multi-factor biometric authentication scheme using bio-hash function, *PLoS One*, vol. 12, no. 5, pp. e0176250, 2017. <https://doi.org/10.1371/journal.pone.0176250>.
- [39] W. Li, Y. Shen, and P. Wang, Breaking Three Remote User Authentication Systems for Mobile Devices, *J. Signal Process. Syst.* vol. 90, no. 8, pp. 1179-1190, 2018. <https://doi.org/10.1007/s11265-017-1305-z>.

A Review on Bio-inspired Optimization Method for Supervised Feature Selection

Montha Petwan¹

Faculty of Science and Technology
Suratthani Rajabhat University
Khun Taleay, Muang, Surat Thani, 84100, Thailand

Ku Ruhana Ku-Mahamud²

School of Computing
Universiti Utara Malaysia, 06010 Sintok, Kedah, Malaysia
Shibaura Institute of Technology, Tokyo, Japan

Abstract—Feature selection is a technique that is commonly used to prepare particular significant features or produce understandable data for improving the task of classification. Bio-inspired optimization algorithms have been successfully used to perform feature selection techniques. The exploration and exploitation mechanism that is based on the inspiration of living things to find a food source and the biological evolution in nature. Nevertheless, irrelevant, noisy, and redundant features persist from the situation of fall into local optima in case of high dimensionality. Thus, this review is conducted to shed some light on techniques that have been used to overcome the problem. The taxonomy of bio-inspired algorithms is presented, along with its performances and limitations, followed by the techniques used in supervised feature selection in term of data perspectives and applications. This review paper has also included the analysis of supervised feature selection on large dataset which showed that recent studies focus on metaheuristic methods because of their promising results. In addition, a discussion of some open issues is presented for further research.

Keywords—Bio-inspired optimization; swarm intelligence; evolutionary algorithm; machine learning

I. INTRODUCTION

Application of science and engineering such as image classification, machine learning, text mining, image retrieval, intrusion detection, and biology analysis, containing huge number of features that are used for information processing and decision making [1],[2],[3],[4]. These applications must be approached carefully due to the abundance of data dimensions, described as the term of big data [5],[6]. Big data has well-known properties such as velocity, variety, value, volume, and veracity [7]. Memory space and costs will increase with escalated data volume. The variety of data makes data integration difficult due to data structured differently. The veracity is a noisy data and quality fluctuation of data when acquired from multi-sources [8],[9]. In solving this problem, irrelevant and redundant feature elements have to be eliminated from superfluous features leaving only data that represent the actual meaning of all features. This complication can be resolved by utilizing feature selection methods. Feature selection techniques are envisioned in the data-preprocessing stage to reduce the dimensionality by selecting the significant features from the original features in problem domain with higher performance of the task as well as speeding up the algorithm [10],[11],[12].

Filter methods such as the Fisher score [13] rank each feature independently under the Fisher criterion in a supervised model that successfully reduces the feature's size. However, this technique cannot determine the correlation among different features. The permutation of individual features does not necessarily and cannot achieve the desired feature set. Thus, the subset of feature is suboptimal [14]. Linear discriminant analysis is one traditional technique to enhance selected features by maximizing the proportion between the intraclass distribution and interclass distributions. Meanwhile, the inverse matrix calculation within-class distribution of linear discriminant analysis will be tolerated when dealing with a small number of labeled data [15]. Wrapper methods depend on a certain classification algorithm in evaluating the selected feature [16]. Hybrid methods attempt in incorporating the dominant characteristic of filter and wrapper models. The aim of filter phase is to reduce the feature dimensionality. The wrapper stage is then used to select the most optimal feature subset [17]. The author in [18] proposed a filter-wrapper algorithm which applied minimum redundancy maximum relevance algorithm to carry out a local search mechanism. Rough set theory and conditional entropy algorithm as filter method introduced in [19] were proposed in selecting the most significant features from a whole set as the initial population. The wrapper approach, which employs the k-nearest neighbor (KNN) algorithm, was then used as an evaluator of their quality of feature combination. The wrapper method can obtain a higher accuracy rate because they determine the correlations in each feature. In fact, the hybrid method achieves high accuracy with respect to the characteristic of wrapper and filter. However, wrappers are computationally more expensive, have less generalization than the filter and their performances are highly dependent on the particular classifier.

Searching and optimization methods are used to find the best solution of many classification problems. Recently, such techniques exhibit their capacity in dealing with Non-Polynomial (NP) hard problems. Finding the most significant features within reasonable amount of time is also considered as NP-hard problem [20], [21]. For example, there are N features containing in the dataset, 2N features are to be generated and evaluated, which will increase the computational cost, especially when each subset is executed by using the wrapper method. As a consequence finding the possible subsets using the exhaustive search or best-first search technique is not a great choice. Recently, an increasing number of optimization approaches have been focused in handling the issue of both

numerical and combinatorial optimization. Solutions have proposed optimal feature subsets and intensively developing based on a variety of metaheuristic methods. An optimization problem focuses on finding the optimal value which corresponds to the maximizing or minimizing one of its performance criteria, or multiple objectives have been proposed. Metaheuristics search strategy with a population-based approach has shown attractive competency in coping with the different character of optimization problem scenarios that can be used to handle the feature selection tasks [9],[22].

Metaheuristic optimization methods are getting inspiration evidentially from nature. Its mechanism and capabilities are extraordinarily magical and mysterious that researchers have focused on mapping the natural phenomenon onto intelligence algorithms. For example, finding the food source for the ant by using the shortest path through indirect communication with each other; interaction between organisms to fully matured human being; balancing the ecosystem; hunting movement and echolocation mechanism. Their abilities have been described to solve the complicated problem independently from elementary initial populations and parameters with little or no knowledge of the feature space. Thus, every feature and natural phenomenon used a suitable strategy in getting the best solution. This approach was able to find an optimal solution although simple optimization strategy has been used. One of the dominant categories of metaheuristics optimization methods is bio-inspired optimization. The bio-inspired optimization impersonates the various natural creatures behaviors like fishes, insects, bird swarms, terrestrial animals, reptiles, humans, and other phenomena [5]. The bio-inspired optimization family has emerged and applied in a proposal of feature selection applications, for instance, text mining, information retrieval, robotics, network security, biomedical engineering, power systems, business, agriculture and many more. Their behavior is a random decision that categorizes them as a randomized algorithm. Formulating a bio-inspired optimization algorithm involves modeling a proper problem representation, calculation the obtained solution efficiency through fitness evaluator and identifying operators to generate a new set of solutions [23]. Though, as previously stated, authors have divided a prevalent group in these approaches based on the evolutionary of biological paradigms.

Summarizations of feature selection algorithms have been performed by [11],[24],[25]. These studies have focused on the subset generation techniques in certain application, feature selection by using swarm-based algorithms together with particular classification and clustering tasks. However, a comprehensive overview of supervised feature subset selection based on bio-inspired optimization algorithms in obtaining the most optimal subset association with data perspectives and different applications was not performed. This paper proposes the bio-inspired optimization algorithm taxonomy according to natural biological inspiration as well as the areas in which these algorithms have been employed.

This paper is organized as follows: in Section 2, the description of the taxonomy of the bio-inspired optimization algorithms. The supervised feature selection using bio-inspired optimization algorithms is presented in Section 3. The analysis of technique for large datasets is introduced in Section 4 and

discussion of the supervised feature subsets selection using bio-inspired optimization algorithms are provided in Section 5. Finally, in Section 6, the conclusions are presented together with further research directions in supervised feature subset selection.

II. TAXONOMY OF BIO-INSPIRED OPTIMIZATION ALGORITHM

Bio-inspired optimization algorithms have been identified as an excellent approach and play an essential part in finding the best optimum solution in different problem domains. This type of algorithms imitated from the systematic behaviors of natural biological evolutionary such as mutation, selection along with distributed collective of living organisms, including birds, ants and wild animal. These algorithms exhibit high level of diversity, robustness, dynamic, simplicity, and fascinating phenomena as comparison with other existing methods. Studies in computer science area have been broadly used the various bio-inspired optimization algorithms in many kind of literature, like looking for the optimal solutions for hard and complexity problem domains [26],[27],[28]. The popular salient and accomplish classes or directions in bio-inspired optimization algorithms that mimic the biological collective behavior of animals, biophysical environment, and cooperation between species respectively [5],[29],[30]. This review paper aims to form the optimization algorithms category according to the area of such inspiration to perform a widespread view over the domain. This paper attempts to categorize the sources of bio-inspiration into swarm intelligence algorithms, evolutionary-based algorithms, and ecology-based algorithms [5],[31],[32]. Nature inspired algorithm consists of bio-inspired and physics-based algorithms. Swarm intelligence, evolutionary ecology-based algorithms are all bio-inspired algorithms. These algorithms are global optimization metaheuristics that search for solution in stochastic scheme within an appropriately runtime procedure. Generally, they start with initial solution and then generate the best solution in the next iteration. The important mechanism of metaheuristic algorithm is balancing between exploitation and exploration that may produce global optimal solution. Exploitation mechanism (intensification) contributes to the agent convergence in optimality. Exploration mechanism (diversification) prevents the loss of diversity which occurs when algorithm get trapped in local optima [33]. Bio-inspired algorithms help to tackle the global optimization problem for selecting feature subsets in the classification area by extracting and exploiting the collective local and global behavior schemes. Studies have now focused on increasing the performance of the search competency in the problem space and efficiently selecting a minimal and discriminating feature subset.

A. Swarm Intelligence Algorithm

Swarm intelligence is defined as an interactive system with the multi-agent. The system is the emergent of intelligent behavior collaborating with the collecting to complete the particular objective that cannot be completed by a single agent or acting alone [34],[35],[36]. Two expressive behaviors of swarm-based system consist of the self-organization and decentralization mechanism of animal living in nature. Self-organization can be identified as state transition rule and

stability through positive and negative feedback. Decentralization system can be described as the collaboration in groups through the state of environment to collect the communication. The examples of emergent swarm behavior in nature are bat echolocation, birds flocking, fish school, bee mating, mosquitoes host-seeking, cockroaches' infestation, ant foraging, sea creatures, and many others. Swarm intelligence approach has remarkable results in solving a wide range of NP-hard problems and really becomes nearly practical in the different real world problem domains. Furthermore, the number of possible solutions has gone up significantly in the problem that frequently leads to be indefinite. Swarm intelligence is used to solve real-world nonlinear problem applications considering many applications of sciences, engineering, data mining, machine learning, computational intelligence, business and marketing, bioinformatics, and industries. This paper emphasizes different swarm-based algorithms entirely starting from the behavioral and biological creatures perspective, which are specified in the life cycle of the insects, birds and animals (amphibians and mammals) [37].

According to [36],[38] used the term swarm intelligence to describe a system comprised of autonomous robots cooperating to fulfil the task under study. This mechanism only enables to handle with partial and noisy information about their environment, to force with uncertain situations, and to search solutions to complex problems. In this way, many existing theoretical frameworks and algorithms mimic the miracle ability of swarm behavior to deal with different scenario in feature selection problems. Among them, the well-known widespread used are ant colony optimization (ACO) [39], artificial bee colony (ABC) [40], particle swarm optimization (PSO) [41], cuckoo search (CS) [42] and firefly algorithm (FA) [43]. Another popular swarm intelligence technique is the monkey algorithm [44], wolf pack algorithm [45], bee collecting pollen algorithm [46], dolphin partner optimization [47], bat-inspired algorithm (BA) [48], and Hunting Search [49]. Salp swarm algorithm (SSA) has been extensively adapted bio-inspired algorithm on account of its advantage such as: (1) a novelty algorithm, (2) unsophisticated, (3) lesser parameters, and (4) low computation time [50]. Swarm-based approaches are less complex procedure with several parameters to strong solution as compared with evolutionary-based components such as selection, crossover, and mutation.

B. Evolutionary-based Algorithms

Evolutionary algorithm or evolutionary computation is a search method that simulates the biological perspective of generation stands on iterative framework of fittest population selection namely reproduction, mutation, recombination, and selection. This search has taken advantage of the assortment to find the suitable solution by using the historical data that leads to a better new solution. This algorithm simulates Charles Darwin's law of nature evolutionary of "survival of the fittest" in selection process in such environment. Evolutionary algorithms have been designed to search the optimal or near-optimal solution in various optimization frameworks whereas typical statistical techniques may produce ineffective results. The performance of evolutionary algorithms is generally depending on the evolutionary setting. For example, the methods for changing the values from reproduction and

mutation for creating the new populations may yield different optimization results and speeds of convergence. Some of the well-known evolutionary-based approaches are a genetic algorithm (GA), genetic programming (GP), evolution strategy (ES) [51], evolutionary algorithm, and artificial immune system.

C. Ecology-based Algorithm

Ecology-based algorithm has been presented for cooperative stochastic search algorithms. The algorithm mimics an ecosystem balancing on the earth. It relies on the population relationship of individuals in a particular ecosystem. Each population is related to adaptation or optimization strategy in the unit space. This algorithm, like other metaheuristic optimization algorithms such as ACO, ABC, GA, and others, seeks global optimization solutions [52]. The search performance of the individuals in each population are interpreted based on the exploration and exploitation mechanism, and the initial parameters [53]. The ecology-based algorithm is inspired by the ecological concepts of habitation, relationship and interaction, and inheritance ecologically. The examples of well-known ecology-based algorithms in the computer science are biogeography-based optimization, inspiration from the immigration scheme of species or animals to find new environment properly [54],[55]. Flower pollination algorithm (FPA) is presented by Yang [56]. The algorithm is inspired by flowers pollination process with biotic and abiotic pollination forms. Biotic pollination is typically linked with pollinators' livelihood such as birds, bees, bats or insects to transfer the pollen from one to another. Abiotic pollination is a process that depending on wind, rain or water. Biotic and pollination of a different plant are a process of global pollination performing by the Lévy distribution. Abiotic and self-pollination can be described as a local pollination procedure.

III. SUPERVISED FEATURE SELECTION METHOD

The classification task is associated with class labeled and can be classified into three frameworks: supervised [57], unsupervised [58], and semi-supervised [59]. Class labeled participation is based on supervised feature subset selection [60],[61]. On the contrary, unsupervised feature subset selection is a primary challenge due to the class is unlabeled. Meanwhile, semi-supervised feature subset selection approaches utilized both labeled and unlabeled classes [62].

Supervised feature selection has an intention on classification or regression problems. It determines the solution of feature subsets that aim to distinguish the instances from predicted categories or predicting the potential target. After splitting the feature subset selected by supervised feature selection into learning and testing sets, the learning set is learned and evaluated by certain classifiers or regression model. The relevance between feature and classes labels is evaluated via its correlation. Choosing a strategy with a filter method can be independent of the classifier algorithm. In contrast, the wrapper method can take the advantage of the classification or regression performance to evaluate the fitness of selected features from the original set or make utilization the intrinsic predictability of a classification algorithm in embedding feature selection algorithm into their specific

fundamental learning model called embedded method. Finally, the subset features with the unseen data in the test set are employed to label for the result of predefined class or regression target [25]. In the present paper, the proposed supervised feature subset selection methods in improving the classification performance are addressed.

A total of 46 publications on bio-inspired optimization algorithms focusing on supervised feature subset selection from 2018 to 2021 have been reviewed. The papers were obtained from Scopus database in December 2021. These papers might not provide the entire studies, but presents the trend in general. Table I displays the retrieval articles in two broad perspectives: static and dynamic data. The data perspectives are further categorized into two classes: using stand-alone metaheuristic or combination with other metaheuristics (hybridization). The most frequently used algorithms are swarm intelligence group, categorized as insect and reptiles like ACO, ABC and DA have been extremely applied to solve the supervised feature subset selection. In bird group of swarm intelligence, PSO, CS and harris hawk optimization (HHO) are used to combat feature selection problem. In Fig. 1, the overall number of those papers are also illustrated according to the year of publication. Research on swarm intelligence has significantly increased in 2019.

However, greater interest has shown on swarm-based algorithm such as PSO, ACO, ABC and FA as compared to evolutionary-based and ecology-based. This paper reveals that swarm-based algorithms have advantage in controlling their behavior autonomously, self-organization and adaptability [5],[34]. Moreover, the ecology-based algorithm has emerged as a new algorithm to show the performance for the supervised feature selection area. These algorithms with different evaluation methods (filter, wrapper, and hybrid) are described in this following subsection.

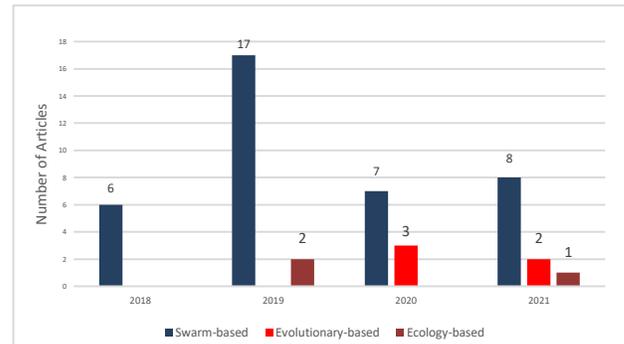


Fig. 1. Usage of the Bio-inspired Algorithm Applied in Supervised Feature Selection.

TABLE I. SUPERVISED FEATURE SELECTION ALGORITHM WITH DIFFERENT DATA PERSPECTIVES

Inspiration	Data perspective				Total
	Static data		Dynamic data		
	Stand alone	Hybridization	Stand alone	Hybridization	
Swarm intelligence Insects and reptiles	[17], [21], [63], [64], [65], [66], [67], [68], [69], [70]	[19], [71]	[72]	[73]	14
Birds	[3], [10], [26], [74], [75], [76]	[27]	[77], [78], [79]	-	10
Terrestrial animals	[28], [80]	-	-	[81], [82], [83]	5
Sea creatures	[1], [18], [84], [85], [86], [87]	[33]	[88]	[89]	9
Evolutionary	[90], [91]	[92]	[93], [94]	-	5
Ecology-based	[95], [96]	[97]	-	-	3
Total	28	6	7	5	46

A. Filter Methods

Filter methods are generally less computational complexity than wrapper approaches. Filter methods evaluate characteristics of data based on some predefined criteria instead of using capability of certain learning algorithm. In this method, the evaluated features that have lower ranking criteria are filtered out. The filter method can be generally broken down into univariate and multivariate schemes. Individual features are ranked in univariate scheme, while the multivariate scheme ranked each feature simultaneously. Many studies have used different evaluation measure techniques to enhance the accomplishment of feature subset selection [25]. These techniques are relief algorithm [98], feature correlation [99], mutual information [100],[101], Fisher score [102] and principal component analysis (PCA) [103]. These techniques are not guided by a certain learning algorithm, such features could be deteriorated the decision making procedures. Requirement efficient metaheuristic methods can improve the performance when it carries on with large-scale feature space.

In [68], ten chaotic maps have been employed in searching process of the dragonfly algorithm (DA) for choosing the optimal extracted features for achieving convergence speed and efficiency of toxicity drug identification task. The selected features from chaotic DA were then fed into a support vector machine (SVM). The experiment indicated that Gauss chaotic map provide the best performance of DA. In [79] the population initialization in PSO is identified by assigning the Relief scores to distinguish ability of biological features. This technique examined the difference between selected sample as well as its homogeneous and heterogeneous neighbor sample. Then, the threshold selection is used to determine the number of features. The author in [78] carried out streaming features from big datasets through parallel processing with the MapReduce technique, dividing the incoming data into subsection. BA is applied in reduced data dimensionality. Then, the ensemble method with multi-layer perceptron artificial neural network classifier is employed to classify the selected significant features. The result of this work has

shortened the processing time but enhanced the accuracy. The author in [93] fused detection mechanism to scan the existing and incoming of feature drifts. The proposed multi-objective feature selection utilized measurement the quality of the solutions based on mutual information method and GA. The GA is started for evaluating the solutions based on merging population, sorting, and crowding distance mechanism. In addition, [28] proposed reducing the dimension of the binary search space of different scale datasets based on social spider algorithm (SSA), S-shaped and V-shaped transfer function are used to evaluate the binary search space. Each possible best solution is improved to the quality solution through crossover mechanism. The performance of this algorithm named BinSSA4 with crossover operator is superior in terms of fitness values, standard deviation values, number of the selected feature, and the accuracy. In application of network intrusion detection, [66] proposed the performance of ACO by enhancement the exploration process. The proposed algorithm keeps away from falling into a local optimum by designing fitness function, pheromone mitigation and increasing for some special trail.

B. Wrapper Approaches

A wrapper approaches is generally composed of two main steps: 1) explores for a potential feature subset and 2) measure the influential quality of selected features. The model iteratively generates both steps until the predefined stopping condition is met. The possible features are first generated as a feature subset, and then these features are evaluated to measure their quality based on a particular learning algorithm. In other words, the wrapper-based feature selection produces repeatedly until the desired accuracy rate is achieved or the desired minimum number of selected features is acquired and then returned as optimal selected features for a particular problem. However, the well-known impractical condition of wrapper methods is dealing with the high dimensionality of d features ($2d$) or the large scale of the search space. Additionally, this method will be met high complexity running time compared with the filter methods. However, there is a wealth of literatures on wrapper-based feature selection when handling high dimensionality datasets. In [63] the continuous version of original butterfly optimization algorithm (BOA) is performed by S-shaped and V-shaped activation function. The results shown that the S-shaped is able to boost the capacity of original BOA that can achieved better accuracy rate and number of selected features. The study of [69] proposed the grasshopper optimization algorithm (GOA) to obtain the most optimal solution by more repulsion in unexplored search space. Promising regions were exploited by intensity and length scale of attractive function. The literature supplemented the algorithm by using SVM during iterations to deal with unexplored feature space and duplicated features in the selected subset. In [64], the integration between ABC and gradient boosting decision tree algorithm is established to explore the best final result. The initial problem space is spanned based on gradient boosting decision tree to categorize the sample into positive or negative patterns. The performance of ABC algorithm can remove low correlation into reduced preliminary input of those decision algorithm. The study of [71] used differential evolution to perform the pheromone updating rule mechanism of max-min ant system (MMAS) by disrupting the

pheromone deposition over the features space and raising the behavior of ants in exploring for optimal feature subset for the benefit of classification task. The binary version of DA with dynamic behavior transfer functions, S-shaped and V-shaped are incorporated for beneficial better solution from unexplored regions [67]. Different chaotic map is experimented due to the problem of slow convergence speed and getting stuck in local optimum of SSA algorithm. The five chaotic variables are adapted for salp position. The results of this algorithm is compared with original SSA, GA and PSO that outperformed such algorithms [84]. Binary version of integrated grey wolf optimization algorithm (GWO) and PSO have been proposed in [81] to cope feature subset selection. This combining, the velocity and position have been controlled by weighting function to balance the diversification and intensification of proposed algorithm. KNN algorithm with euclidean distance measurement is employed in the wrapper-based method. Two binary forms of whale optimization algorithm (WOA) algorithm have been integrated with evolutionary operators to perform the exploration and exploitation mechanism in seeking the optimal selected feature for increasing classification targets. In the search process, the Tournament and Roulette Wheel Selection mechanisms are used. Crossover and mutation operators are applied to increase the exploitation mechanism of the WOA algorithm. The results showed that WOA with crossover and mutation outperformed GA, PSO and ant lion optimizer [33]. The study of [70] propose a binary FA with two objectives, accuracy rate and reduction rate to reduce the number of features. In this literature, the new formula is proposed by calculating the distance between two fireflies to enhance the quality of exploration and exploitation of search space. The results of algorithm outperformed the PSO. The discrete cosine transform (DCT) with fixed-size window technique was applied in [77] to exploit the current informative features of data streaming as a baseline. Then, the efficient feature subset produced from the PSO algorithm is fed into the KNN classification algorithm for decision processing. The experimental demonstrated the DCT without and with feature selection. The result show that the automatic feature selection process searches the best feature subset that can give higher performance. The combination metaheuristic approach is proposed by using GWO and WOA to enhance a wrapper-based feature subset selection technique. The hybridization is accomplished by improving the mechanism of both algorithms including immature convergence and stagnation to local optima [83]. As sine-cosine algorithm which can be supplemented the exploration stage, [74] combined this algorithm into HHO in exploration phase, as the result effectiveness in exploitation phase can get the quality information. Additionally, the delta factor is injected in exploitation phase. The results of this proposed outperform sine-cosine algorithm and original HHO algorithm for ten datasets out of sixteen in terms of fitness. In terms of accuracy, the proposed outperform other optimization method on eleven datasets out of sixteen. The author in [75] proposed the opposite point exploration and disruption operation due to problem of struggling in local optimization of CS algorithm, this prevents over random search of features. These enhancements improve the exploration phase and feature selection in complex data. The results show that this algorithm can find the maximizing the classification accuracy rate and

reducing the number of features; however, the computational time still increases. In the same manner, [86] has been adapted opposition-based learning (OBL) technique in slime mould algorithm (SMA) to overcome premature convergence and slow movement. Moreover, [89] presented the concept of OBL. The procedure is the WOA run first and at the same time during the run, population is changed by the OBL. To increase the accuracy and speed convergence, it is used as the initial population of FPA. The best feature set from transformed dataset is proposed in [85]. The continuous values are converted into binary search space using empirical threshold δ . This work shows very excellent results when combined with PCA and independent component analysis. As the results, feature selection method only reduces the unnecessary and unimportant features, not the correlated and higher-order dependencies among features. Due to the problem of CS that is over randomly causes blindly preservation the quality solution, the work of [76] applied chaotic map to enhance the exploration mechanism of CS. The proposed two-population elite preservation strategy can find the attractive one of each generation and preserve it. Levy flight is developed to update the position of a cuckoo, and the proposed uniform mutation strategy avoids the trouble that the search space is too large for the convergence of the algorithm due to Levy flight and improves the algorithm exploration ability. The author in [27] proposed binary PSO with FPA, PSO performs as a global search and FPA conducts a fine-tuned search. The study of [10] performed variant of PSO, competitive swarm optimizer with KNN to handle the large scale optimization problem.

Moreover, an ant colony-based approach has been presented to explore the suitable signals feature subset for application of power quality disturbances classification. S-transform fused with time-time transform is employed for detection and feature extraction. The proposed presented the results of classification with and without feature selection [65]. The author in [90] designed a OBL mechanism to both exploration and exploitation in differential evolution variant have been proposed in the application problem of engineering. The diversity measurement is designed to recognize the convergence behavior of OBL variants. Secondly, the explorative opposition and exploitative opposition are distinguished according to the convergence behavior of OBL variants. Finally, the protective mechanism is introduced to obtain a good ratio between exploration and exploitation for better performance without extra fitness evaluations. This literature carried out experiments on the IEEE Congress of Evolutionary Computation (CEC) 2017, CEC 2011 and CEC 2020 test suites that shows superior performance. The problem of over fitting classification in the swine breed has been solved in [95] by applying feature selection technique to reduce many large original features into the most significant porcine single nucleotide polymorphism. Binary FPA is combined with information gain along with the cut-off-point-finding threshold to identify a 0 or 1 value in feature vector and GA bit-flip mutation operator. The result of this study revealed that the proposed technique outperformed the PSO, CS and FA in terms of classification accuracy. The author in [91] proposed three-dimensional reduction of feature space mechanism under deletion conceptual the unimportant features utilizing the feedback information from evolution algorithm such as DE,

GA, PSO. The assistance of dimension reduction mechanism with evolutionary algorithm is effective way in finding a feature subset with higher classification accuracy and smaller number of features. The WOA with SVM is presented for the task of spam recognition in various languages. The model was proposed to perform automatic detection of arrival spams and gives an insight into the most influential features during the detection process [88]. A artificial fish swarm optimization algorithm with a crossover operation have modified for application of text categorization [1]. The method includes the best fish of swarm can be brought together to improve the capacity of local search. To reduce the run time complexity of ACO algorithm, the study of [21], the degree based graph representation of ACO algorithm for the field of speech processing domain have been proposed. The proposed method will have benefit over fully connected graph and contributed more flexibility on the problem space compared to binary connected graph representation.

In addition, FPA is used to propose the elimination of irrelevant features in biomedical data analysis [96]. The diversity of the population and search performance of the FPA algorithm have been increased by adopting the absolute balance group strategy and adaptive Gaussian mutation. The classification rate is evaluated by using KNN. The experimental result reveals that the proposed method outperforms other state-of-the-art methods. The author in [97] proposed hybrid version of biogeography-based optimization and GA in breast density classification. The authors of [73] combined ABC and CS for reduction the number of features to utilize the incoming anomalies detection in network. The binary pigeon inspired optimizer based on cosine similarity concept have been proposed to build optimal solution by calculating the velocity of the pigeons [3]. Another swarm algorithm for intrusion detection was proposed by [82]. The proposed algorithm utilized grey wolf optimization algorithm for selection the most optimal features by controlling the balancing between exploration and exploitation mechanism.

C. Hybrid Methods

Hybrid method aggregates the dominant of multiple feature selection approaches such as filter-wrapper. The principal intention is to increase the stability of the solution by combination salient of different available feature subset selection algorithms. For example, dealing with a small dataset with high dimensionality and dealing with a small combination on the training set will result in completely different solutions. By combining multiple selected features based on different approaches, the solutions are more straightforward, so the quality of the selected features can be preserved. In hybrid methods that is stimulated by the way of bio-inspired algorithm, [17] introduced the ranked informative features using filter method. Their quality obtained feature sets are then evaluated based on wrapper method. Improved memory to keep the best ant and normalized pheromone updating mechanism have also been proposed to enhance the feature selection. A logistic map sequence has been used to perform diversity in the problem space of PSO with a spiral-shaped mechanism integrated as an operation of local search around the known optimal solution boundary. The current position and flying velocity are incorporated with two dynamic correction

factors to improve the exploration and exploitation in the feature space [26]. In [87] applied the Pearson's correlation coefficient and correlation distance to adapt the weights of unrelated and consistent features based on filter manner. The random population WOA functions act as a wrapper algorithm. The results achieve the highest classification accuracy in all datasets, but not shortest length of feature set compared with other algorithms. In this manner, chaotic WOA is introduced [18]. Besides, [92] established integration the strengths of GA and PSO to conquer the tradeoff between exploitation and exploration procedure. The wrapper proposed utilizes artificial neural network to evaluate the fitness of feature set. This method applied the data which pull out from smaller datasets to reduce the processing time of subset selection on high volume datasets. Ant lion optimizer is integrated with hill-climbing technique to find the best solution in work of [19].

Furthermore, the feature selection for anomalies detection in the network has been proposed in [72] based on the FA. This literature employs mutation-based in filtering features and FA with wrapper-based methods for evaluating selected features to C4.5 and Bayesian Networks based classifiers. The author in [94] proposed filter-based, information gain metric and the sorting mechanism of evolutionary algorithm which is efficient multiobjective optimization algorithm. The dataset is the World Health Organization Director-General's speeches during the COVID-19 pandemic period and Stanford Sentiment Treebank. Due to the abundance of irrelevant and redundant data in microarray datasets, authors in [80] proposed information gain and krill herd algorithm to capture only the important features from the original datasets.

IV. ANALYSIS OF TECHNIQUE FOR HUGE DATASET

The goals of proficient data analysis rely on the providing a large amount of data and these purposes is indispensable to deal with analytics on huge datasets. Working toward data analysis, the facilitation of massive data preservation technologies, the revolution of digital technologies, where huge sizes of data are generated with ever-increasing volumes of transactions over time and diversity. The term "huge dataset" was motivated and officially reported in international conference [104]. One of distinctiveness of huge datasets is that the sample size is generated greater than petabytes level and moves very fast when used to describe a given sample. Recently, the modernizations of technologies and Smartphones have enabled users to use online media to communicate with others in a one-way or two-way manner. Some of examples of sources with huge datasets include Facebook, Twitter, blogs, flick, LinkedIn, Pinterest, sensing devices, and web-based email. Huge collections of these data platforms typically have complex structures which received from multiple sources. The information collected from this online source consists of different data formats such as text, image, videos, log, and so on. It is extremely challenging to obtain important hidden data from these social assemblages in an appropriate manner. The increasing rate of digital adoption has been observable since the initiation of the coronavirus, covid-19 pandemic since 2020. The rate of digital growth has increased dramatically, and more are expected in the second half of 2021 as shown in Fig. 2. It shows that the power of social media continues to drive activities to be connected all over the world, with an

increasing number of social media users worldwide, and an impressive step is rapidly approaching. Also, Facebook and Instagram have attained 5.1 billion users. Twitter generated 8 terabytes of data per day, or 80 million tweets per day [105], and Line and Whatsapp generates approximately 2.5 petabytes of data per hour.

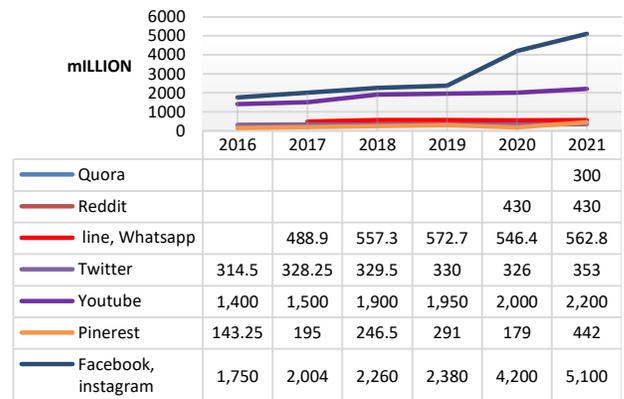


Fig. 2. The Rate of Data Generation in Digital World from 2016 to 2021.

An efficient way in extraction the important information can be measured by the complexity of computing methods and algorithms. Traditional methods and tools are used to operate small and structured data by trail-and-error. This endeavors analysis is not suitable when data sets are large and miscellaneous. This large dataset requires a large amount of memory to store and takes hours or days in the case of using traditional processing methods. At present, a superfluity of different improved techniques are being developed to treat with large amounts of unstructured data sets [106]. The nature of large amounts of data is adequate for learning models to be applied to real-world scenarios efficiently. In addition, the models can be enhanced to digest as much target discrimination as possible. Face forgery detection is an example of an application domain and technique that makes use of a large dataset, with data that is ten times larger than the previous forensic dataset. The increased data source is distorted through face swapping method for robustness the head poses variation due to videos on the internet usually have limitation gesture of head. The feature learning with dimensionality reduction via autoencoder is considered for forensics. The experiment was carried out with as many real-world perturbations as possible. The accuracy results remained low due to the poor quality of the learning set and augmentation method for face diversity [107]. Moreover, the novel chance constrained problem domain is formulated using a huge dataset. A weighted feature reduction operation is proposed to describe a relaxation problem of chance constrained problems. Also, a DE has been adapted and integrated with a pruning technique to force the relaxation concern of chance constrained [108].

In the natural language classification application domain, in [109] employed rule-based classification and Apriori algorithm because it cannot maintain the capability between accuracy and interpretability in the non-big data environment. This problem has been facilitated by proposing the probability integral transform theorem, rule induction and rule selection based on evolutionary optimization. The experimental results show that

fuzzy model-generated models are significantly simpler in terms of classification rate, complexity, and time consumption. A distributed fuzzy decision trees have been proposed due to the problem of time constraints and space requirements. This paper proposed using the MapReduce framework to partition large scale data into binary and multiple decision ways. The relevant features are derived by using information gain method, which will be used in the decision nodes. The author implemented the fuzzy decision tree learning scheme on the Apache Spark framework [110]. Among the given domains that use big data technologies is vehicular ad hoc network to handle their big size data. The author demonstrates a method that entail to detect accidental or irregularity on the way and estimate the distance and time spending on each route in the form of real time system, which allows the user to gain a database having the estimated time spending in all sections, this will serve the vehicles for the reasonable estimated time of attainment consistently throughout their travel and optimize the best route to reach the destination. The experiment reveals that this method effectively warns for crowded vehicles or portions the vehicle overwhelming in all roads, and it can also be used to save road safety [111]. Furthermore, human disease application used huge dataset to prevent the spread of infectious virus to human as the outbreak of covid-19 virus. The epidemic has forced everyone to stay and work at home that has affected on the people's mental health around the world. The global covid-19 pandemic requires a broader overview of data set to analyze the problem of human-to-human transmission of the virus across the country [112]. The covid-19 tracker with HPCC system is used to predict the future trend of the covid-19 outbreak, but this tool is limited in such a way that it cannot explain other factors that may influence trends such as mobility, local weather conditions, and so on. The smoothing filter of the covid-19 tracker is capable of eliminating irregular data transfers, but the system is still incomplete, with only a minor effect on the natural time vector. The automatically ingested data was pulled into the system for data cleaning and then extraction of the important data for subsequent analysis. The system will also be automatically executed when the new information enters to the system [113]. Additionally, human mobilities are extremely restricted affected by the covid-19 situation. Autonomous robot systems are becoming very important and can be conducted to replace human service work such as serving medicine to the infectious patients. The first large-scale elevator panel dataset has been made public in order to challenge the problem of inter-floor navigation. The deep learning based is used to recognize autonomous elevator operation. The performance of that model and dataset is compared to popular network such as ResNet, PSPNet and U-Net. The results of ResNet have the best performance when compared to the remaining networks [114].

V. DISCUSSION

The recent advance in feature selection algorithms have grown exponentially across a wide range of application domains. The most explored areas continue to be in fixed data or static such as bioinformatics and image processing. In addition, social media platforms such as Twitter, Facebook, blogs, and wikis are prevalent in streaming data or dynamic. Among the bio-inspired algorithms, the swarm-based

algorithms have been applied in many areas to produce excellent solutions for problems where the characteristic of the problem is NP-hard, which otherwise produce sub-optimum solutions and consume a lot of processing power. Due to the scalability and simplicity of SI-based algorithms, they have become the first choice in producing the outcomes for an optimization problem. The present researchers expect this review paper to provide other researchers working on different bio-inspired optimization algorithms to effectively and efficiently handle new challenges in supervised feature subset selection.

To this day, many valuable feature selection algorithms have been extensively developed for real-world application and theoretical analysis. However, the present researchers believe that more intelligent behavior from nature can be applied to improve the solutions in this field. There are several contributions and issues related to feature selection method. Firstly, according to the enormous increment in the volume of the data, the recent feature subset selection algorithms may be threatened especially in terms of scalability with online datasets. Secondly, the performance of supervised feature subset selection algorithms is commonly evaluated by the compromised accuracy. As a result, algorithms adaptation should be an essential concern when exploring new searching space and exploitation of the best solution that selected in each iteration to gain most optimal feature. It is determined as the affectability of a feature subset selection algorithm to feature combination during the training phase. Finally, in statistical feature selection algorithms, feature weighting techniques are frequently used to identify the number of selected features. In this method, the number of optimal selected features is discarded. Furthermore, a large number of selected features will jeopardize the learning performance due to the inclusion of noisy, irrelevant, and redundant features. By extension, it should not use a small number of feature subsets because some relevant features are excluded. In practice, many researchers have usually used a metaheuristic way to find and evaluate the candidate feature subset and feed the number of selected features that have the best classification accuracy, but the whole process suffers from computation time. The challenging problem in this domain is determining the optimal number of selected features. Furthermore, the present researchers believe that the selection of the evaluation criteria is also a crucial aspect which requires deeper investigation.

VI. CONCLUSION

Feature selection techniques aim to provide effective pre-processing of data in eliminating redundant and irrelevant features. It is a fundamental method in preparing the data which is clean and intelligible. It has been an interesting field of research work that has proven to be extremely useful in many application domains including image recognition, machine learning, web and text mining, pattern classification, and medical diagnosis in both offline and online platforms. The past few years the performance of many novel feature selection methods has augmented. This can be observed that, several optimization algorithms have been presented to solve problematic feature selection by optimization the value to gain the best suitable solution. Such algorithms that are inspired by the natural, biological and ecological behavior is to produce

good solutions. However, there are only several bio-inspired algorithms that have been proposed. This study has reviewed works on the enhancement of solving the feature selection problems. In addition, the authors have detailed the algorithms that are practical among which are wolf-based algorithms, salp, and biogeography algorithms. Highlights on each algorithm have been presented, followed by recent advances in the literature and problem domains for each. Nonetheless, it is important to highlight that these algorithms have yet to be demonstrated impressively in large-scale datasets such as streaming data and linked data. This brings up subsequent research application domain such as medical, the environment, and social science.

ACKNOWLEDGMENT

The authors would like to thank the Ministry of Higher Education Malaysia for funding this study under the Transdisciplinary Research Grant Scheme, TRGS/1/2018/UUM/02/3/3 (S/O code 14163).

REFERENCES

- [1] Thiagarajan and N. Shanthi, "A modified multi objective heuristic for effective feature selection in text classification," *Cluster Comput.*, vol. 22, pp. 10625–10635, 2019.
- [2] K. Sankar and P. Uma Maheswari, "A dynamic wrapper-based feature selection for improved precision in content-based image retrieval," *Concurr. Comput.*, 2019.
- [3] H. Alazzam, A. Sharieh, and K. E. Sabri, "A feature selection algorithm for intrusion detection system based on pigeon inspired optimizer," *Expert Syst. Appl.*, vol. 148, p. 113249, 2020.
- [4] L. Ignaczak, G. Goldschmidt, C. A. D. Costa, and R. D. R. Righi, "Text mining in cybersecurity: a systematic literature review," *ACM Comput. Surv.*, vol. 54, no. 7, pp. 1–36, 2021.
- [5] M. Sharma and P. Kaur, "A comprehensive analysis of nature-inspired meta-heuristic techniques for feature selection problem," *Arch. Comput. Methods Eng.*, no. 0123456789, 2020.
- [6] M. Rong, D. Gong, and X. Gao, "Feature selection and its use in big data: challenges, methods, and trends," *IEEE Access*, vol. 7, pp. 19709–19725, 2019.
- [7] B. H. Nguyen, B. Xue, and M. Zhang, "A survey on swarm intelligence approaches to feature selection in data mining," *Swarm Evol. Comput.*, vol. 54, no. April 2019, p. 100663, 2020.
- [8] M. Cherrington, Q. Xu, D. Airehrour, S. Wade, J. Lu, and S. Madanian, "Feature selection methods for linked data: Limitations, capabilities and potentials," in *BDCAT 2019 - Proceedings of the 6th IEEE/ACM International Conference on Big Data Computing, Applications and Technologies*, 2019, pp. 103–112.
- [9] C. Dhaenens and L. Jourdan, "Metaheuristics for data mining: Survey and opportunities for big data," *40R*, vol. 17, no. 2, pp. 115–139, 2019.
- [10] S. Gu, R. Cheng, and Y. Jin, "Feature selection for high-dimensional classification using a competitive swarm optimizer," *Soft Comput.*, vol. 22, no. 3, pp. 811–822, Feb. 2018.
- [11] S. A. K. M. A. M. A. K. Nazish Naheed Muhammad Shaheen, "Importance of features selection, attributes selection, challenges and future directions for medical imaging data: a review," *Comput. Model. Eng. & Sci.*, vol. 125, no. 1, pp. 315–344, 2020.
- [12] S. B. Kotsiantis, "Supervised machine learning: a review of classification techniques," *Informatica*, vol. 31, pp. 249–268, 2007.
- [13] M. Gan and L. Zhang, "Iteratively local fisher score for feature selection," *Appl. Intell.*, vol. 51, no. 8, pp. 6167–6181, 2021.
- [14] Hashemi, M. B. Dowlatshahi, and H. Nezamabadi-pour, "Ensemble of feature selection algorithms: a multi-criteria decision-making approach," *Int. J. Mach. Learn. Cybern.*, vol. 13, no. 1, pp. 49–69, 2022.
- [15] G. Zhao and Z. Zhou, "Efficient linear feature extraction based on large margin nearest neighbor," *IEEE Access*, vol. 7, pp. 78616–78624, 2019.
- [16] M. Kusy and R. Zajdel, "A weighted wrapper approach to feature selection," *Int. J. Appl. Math. Comput. Sci.*, vol. 31, no. 4, pp. 685–696, 2021.
- [17] M. Ghosh, R. Guha, R. Sarkar, and A. Abraham, "A wrapper-filter feature selection technique based on ant colony optimization," *Neural Comput. Appl.*, vol. 3, 2019.
- [18] R. Guha, M. Ghosh, S. Mutsuddi, R. Sarkar, and S. Mirjalili, "Embedded chaotic whale survival algorithm for filter–wrapper feature selection," *Soft Comput.*, vol. 24, no. 17, pp. 12821–12843, 2020.
- [19] M. M. Mafarja and S. Mirjalili, "Hybrid binary ant lion optimizer with rough set and approximate entropy reducts for feature selection," *Soft Comput.*, vol. 23, no. 15, pp. 6249–6265, 2019.
- [20] X. He, M. Ji, C. Zhang, and H. Bao, "A variance minimization criterion to feature selection using laplacian regularization," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 10, pp. 2013–2025, 2011.
- [21] R. R. Rajoo and R. Abdul Salam, "Ant colony optimization based subset feature selection in speech processing: constructing graphs with degree sequences," *Int. J. Adv. Sci. Eng. Inf. Technol.*, vol. 8, no. 4–2, p. 1728, 2018.
- [22] J. Swan et al., "Metaheuristics 'In the Large,'" *Eur. J. Oper. Res.*, vol. 297, no. 2, pp. 393–406, 2022.
- [23] Jain and D. Zongker, "Feature selection: evaluation, application, and small sample performance," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 19, no. 2, pp. 153–158, 1997.
- [24] L. Brezočnik, I. Fister, and V. Podgorelec, "Swarm intelligence algorithms for feature selection: a review," *Appl. Sci.*, vol. 8, no. 9, 2018.
- [25] Telikani, A. Tahmassebi, W. Banzhaf, and A. H. Gandomi, "Evolutionary machine learning: a survey," *ACM Comput. Surv.*, vol. 54, no. 8, pp. 1–35, 2021.
- [26] K. Chen, F. Y. Zhou, and X. F. Yuan, "Hybrid particle swarm optimization with spiral-shaped mechanism for feature selection," *Expert Syst. Appl.*, vol. 128, pp. 140–156, 2019.
- [27] M. A. Tawhid and A. M. Ibrahim, "Hybrid binary particle swarm optimization and flower pollination algorithm based on rough set approach for feature selection problem," in *Nature-Inspired Computation in Data Mining and Machine Learning*, X.-S. Yang and X.-S. He, Eds. Cham: Springer International Publishing, 2020, pp. 249–273.
- [28] BAŞ and E. ÜLKER, "An efficient binary social spider algorithm for feature selection problem," *Expert Syst. Appl.*, vol. 146, p. 113185, 2020.
- [29] R. Diao and Q. Shen, "Nature inspired feature selection meta-heuristics," *Artif. Intell. Rev.*, vol. 44, no. 3, pp. 311–340, 2015.
- [30] B. Xue, M. Zhang, W. N. Browne, and X. Yao, "A survey on Evolutionary Computation approaches to feature selection," *IEEE Trans. Evol. Comput.*, vol. 20, no. 4, pp. 606–626, 2016.
- [31] Dhiman and V. Kumar, "Spotted hyena optimizer: A novel bio-inspired based metaheuristic technique for engineering applications," *Adv. Eng. Softw.*, vol. 114, pp. 48–70, 2017.
- [32] S. Goel, A. Sharma, and V. K. Panchal, "Performance analysis of bio-inspired techniques," in *Proceedings of the Third International Conference on Soft Computing for Problem Solving*, 2014, pp. 831–844.
- [33] M. Mafarja and S. Mirjalili, "Whale optimization approaches for wrapper feature selection," *Appl. Soft Comput.*, vol. 62, pp. 441–453, 2018.
- [34] S. Almufti, "Using swarm intelligence for solving NPhard problems," *Acad. J. Nawroz Univ.*, vol. 6, no. 3, pp. 46–50, 2017.
- [35] Y. Li, "Solving TSP by an ACO-and-BOA-based hybrid algorithm," in *2010 International Conference on Computer Application and System Modeling (ICCASM 2010)*, 2010, vol. 12, pp. V12-189-V12-192.
- [36] R. S. Parpinelli and H. S. Lopes, "New inspirations in swarm intelligence: A survey," *Int. J. Bio-Inspired Comput.*, vol. 3, no. 1, pp. 1–16, 2011.
- [37] K. Sethi, X. Li, L. Cheng, S. Yadavalli, and L. Zhang, "Swarm intelligence: A review of algorithms," no. March, p. 494, 2017.

- [38] Khosravianian, M. Rahmanimanesh, and P. Keshavarzi, "Discrete social spider algorithm for solving traveling salesman problem," *Int. J. Comput. Intell. Appl.*, vol. 20, no. 03, p. 2150020, 2021.
- [39] M. Dorigo and G. Di Caro, "Ant colony optimization: a new metaheuristic," in *Proceedings of the 1999 Congress on Evolutionary Computation-CEC99 (Cat. No. 99TH8406)*, 1999, vol. 2, pp. 1470-1477 Vol. 2.
- [40] D. Karaboga and B. Basturk, "A powerful and efficient algorithm for numerical function optimization: artificial bee colony (ABC) algorithm," *J. Glob. Optim.*, vol. 39, no. 3, pp. 459-471, 2007.
- [41] Kennedy and R. Eberhart, "Particle swarm optimization," in *Proceedings of ICNN'95 - International Conference on Neural Networks*, 1995, vol. 4, pp. 1942-1948 vol.4.
- [42] X. Yang and Suash Deb, "Cuckoo Search via Lévy flights," in *2009 World Congress on Nature Biologically Inspired Computing (NaBIC)*, 2009, pp. 210-214.
- [43] X.-S. Yang, "Firefly Algorithms for Multimodal Optimization," in *Stochastic Algorithms: Foundations and Applications*, 2009, pp. 169-178.
- [44] Mucherino and O. Seref, "Monkey search: A novel metaheuristic search for global optimization," in *AIP Conference Proceedings*, 2007, vol. 953, no. 1, pp. 162-173.
- [45] S. Wu, F. Zhang, and L. Wu, "New swarm intelligence algorithm-wolf pack algorithm," *Syst. Eng. Electron.*, vol. 35, no. 11, pp. 2430-2438, 2013.
- [46] X. Lu and Y. Zhou, "A novel global convergence algorithm: bee collecting pollen algorithm," in *Advanced Intelligent Computing Theories and Applications. With Aspects of Artificial Intelligence*, 2008, pp. 518-525.
- [47] Y. Shiqin, J. Jianjun, and Y. Guangxing, "A Dolphin Partner Optimization," in *2009 WRI Global Congress on Intelligent Systems*, 2009, vol. 1, pp. 124-128.
- [48] X.-S. Yang, "A New Metaheuristic Bat-Inspired Algorithm," in *Nature Inspired Cooperative Strategies for Optimization (NICSO 2010)*, J. R. González, D. A. Pelta, C. Cruz, G. Terrazas, and N. Krasnogor, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010, pp. 65-74.
- [49] R. Oftadeh, M. J. Mahjoob, and M. Shariatpanahi, "A novel metaheuristic optimization algorithm inspired by group hunting of animals: Hunting search," *Comput. Math. with Appl.*, vol. 60, no. 7, pp. 2087-2098, 2010.
- [50] S. Mirjalili, A. H. Gandomi, S. Z. Mirjalili, S. Saremi, H. Faris, and S. M. Mirjalili, "Salp Swarm Algorithm: A bio-inspired optimizer for engineering design problems," *Adv. Eng. Softw.*, vol. 114, pp. 163-191, 2017.
- [51] Zhang, M. Chen, X. Xu, and G. G. Yen, "Multi-objective evolution strategy for multimodal multi-objective optimization," *Appl. Soft Comput.*, vol. 101, p. 107004, 2021.
- [52] S. H. A. Rahmati and M. Zandieh, "A new biogeography-based optimization (BBO) algorithm for the flexible job shop scheduling problem," *Int. J. Adv. Manuf. Technol.*, vol. 58, no. 9-12, pp. 1115-1129, 2012.
- [53] R. Alroomi, F. A. Albasri, and J. H. Talaq, "Essential modifications on biogeography-based optimization algorithm," *Comput. Sci. Inf. Technol.*, pp. 141-160, 2013.
- [54] D. Simon, "Biogeography-Based Optimization," *IEEE Trans. Evol. Comput.*, vol. 12, no. 6, pp. 702-713, Dec. 2008.
- [55] W. Gong, Z. Cai, and C. X. Ling, "DE/BBO: A hybrid differential evolution with biogeography-based optimization for global numerical optimization," *Soft Comput.*, vol. 15, no. 4, pp. 645-665, 2011.
- [56] X.-S. Yang, "Flower pollination algorithm for global optimization," in *Unconventional Computation and Natural Computation*, 2012, pp. 240-249.
- [57] S. B. Kotsiantis, "Feature selection for machine learning classification problems: A recent overview," *Artif. Intell. Rev.*, vol. 42, no. 1, pp. 157-176, 2011.
- [58] Almazini and K. Ku-Mahamud, "Adaptive technique for feature selection in modified graph clustering-based ant colony optimization," *Int. J. Intell. Eng. Syst.*, vol. 14, p. 2021, Apr. 2021.
- [59] R. Sheikhpour, M. A. Sarram, S. Gharaghani, and M. A. Z. Chahooki, "A Survey on semi-supervised feature selection methods," *Pattern Recognit.*, vol. 64, pp. 141-158, 2017.
- [60] Z. Zhao, L. Wang, H. Liu, and J. Ye, "On similarity preserving feature selection," *IEEE Trans. Knowl. Data Eng.*, vol. 25, no. 3, pp. 619-632, Mar. 2013.
- [61] S. Xiang, F. Nie, G. Meng, C. Pan, and C. Zhang, "Discriminative least squares regression for multiclass classification and feature selection," *IEEE Trans. Neural Networks Learn. Syst.*, vol. 23, no. 11, pp. 1738-1754, Nov. 2012.
- [62] Z. Xu, I. King, M. R. Lyu, and R. Jin, "Discriminative semi-supervised feature selection via manifold regularization," *IEEE Trans. Neural Networks*, vol. 21, no. 7, pp. 1033-1047, Jul. 2010.
- [63] S. Arora and P. Anand, "Binary butterfly optimization approaches for feature selection," *Expert Syst. Appl.*, vol. 116, pp. 147-160, 2019.
- [64] Rao et al., "Feature selection based on artificial bee colony and gradient boosting decision tree," *Appl. Soft Comput.*, vol. 74, pp. 634-642, 2019.
- [65] U. Singh and S. N. Singh, "A new optimal feature selection scheme for classification of power quality disturbances based on ant colony framework," *Appl. Soft Comput.*, vol. 74, pp. 216-225, 2019.
- [66] H. Peng, C. Ying, S. Tan, B. Hu, and Z. Sun, "An improved feature selection algorithm based on ant colony optimization," *IEEE Access*, vol. 6, pp. 69203-69209, 2018.
- [67] Mafarja et al., "Binary dragonfly optimization for feature selection using time-varying transfer functions," *Knowledge-Based Syst.*, vol. 161, pp. 185-204, 2018.
- [68] G. I. Sayed, A. Tharwat, and A. E. Hassanien, "Chaotic dragonfly algorithm: an improved metaheuristic algorithm for feature selection," *Appl. Intell.*, vol. 49, no. 1, pp. 188-205, 2019.
- [69] Zakeri and A. Hokmabadi, "Efficient feature selection method using real-valued grasshopper optimization algorithm," *Expert Syst. Appl.*, vol. 119, pp. 61-72, 2019.
- [70] S. Maza and D. Zouache, "Binary firefly algorithm for feature selection in classification," *2019 Int. Conf. Theor. Appl. Asp. Comput. Sci. ICTAACS 2019*, 2019.
- [71] J. M. Montemayor and R. V. Crisostomo, "Feature selection in classification using binary max-min ant system with differential evolution," in *2019 IEEE Congress on Evolutionary Computation (CEC)*, 2019, pp. 2559-2566.
- [72] S. B and M. K., "Firefly algorithm based feature selection for network intrusion detection," *Comput. Secur.*, vol. 81, pp. 148-155, 2019.
- [73] H. S. Al-Safi, Z. I. R. Hani, and M. M. Abdul Zahra, "Using a hybrid algorithm and feature selection for network anomaly intrusion detection," *J. Mech. Eng. Res. Dev.*, vol. 44, no. 4, pp. 253-262, 2021.
- [74] Hussain, N. Neggaz, W. Zhu, and E. H. Houssein, "An efficient hybrid sine-cosine Harris hawks optimization for low and high-dimensional feature selection," *Expert Syst. Appl.*, vol. 176, no. July 2020, p. 114778, 2021.
- [75] kelidari and J. Hamidzadeh, "Feature selection by using chaotic cuckoo optimization algorithm with levy flight, opposition-based learning and disruption operator," *Soft Comput.*, vol. 25, no. 4, pp. 2911-2933, 2021.
- [76] Wang, Y. Gao, J. Li, and X. Wang, "A feature selection method by using chaotic cuckoo search optimization algorithm with elitist preservation and uniform mutation for data classification," *Discret. Dyn. Nat. Soc.*, vol. 2021, 2021.
- [77] Ö. Aydođdu and M. Ekinci, "An approach for streaming data feature extraction based on discrete cosine transform and particle swarm optimization," *Symmetry (Basel)*, vol. 12, no. 2, 2020.
- [78] D. Renuka Devi and S. Sasikala, "Online feature selection (OFS) with accelerated bat algorithm (ABA) and ensemble incremental deep multiple layer perceptron (EIDMLP) for big data streams," *J. Big Data*, vol. 6, no. 1, 2019.
- [79] Y. Xue, W. Jia, and A. X. Liu, "A particle swarm optimization with filter-based population initialization for feature selection," *2019 IEEE Congr. Evol. Comput. CEC 2019 - Proc.*, pp. 1572-1579, 2019.
- [80] G. Zhang, J. Hou, J. Wang, C. Yan, and J. Luo, "Feature selection for microarray data classification using hybrid information gain and a

- modified binary krill herd algorithm,” *Interdiscip. Sci. Comput. Life Sci.*, vol. 12, no. 3, pp. 288–301, 2020.
- [81] Q. Al-Tashi, S. J. Abdul Kadir, H. M. Rais, S. Mirjalili, and H. Alhussian, “Binary optimization using hybrid grey wolf optimization for feature selection,” *IEEE Access*, vol. 7, pp. 39496–39508, 2019.
- [82] H. Almazini and K. Ku-Mahamud, “Grey wolf optimization parameter control for feature selection in anomaly detection,” *Int. J. Intell. Eng. Syst.*, vol. 14, p. 2021, Feb. 2021.
- [83] Mafarja, A. Qasem, A. A. Heidari, I. Aljarah, H. Faris, and S. Mirjalili, “Efficient hybrid nature-inspired binary optimizers for feature selection,” *Cognit. Comput.*, vol. 12, no. 1, pp. 150–175, 2020.
- [84] E. Hegazy, M. A. Makhlof, and G. S. El-Tawel, “Feature selection using chaotic salp swarm algorithm for data classification,” *Arab. J. Sci. Eng.*, vol. 44, no. 4, pp. 3801–3816, 2019.
- [85] S. S. Shekhawat, H. Sharma, S. Kumar, A. Nayyar, and B. Qureshi, “BSSA: Binary salp swarm algorithm with hybrid data transformation for feature selection,” *IEEE Access*, vol. 9, pp. 14867–14882, 2021.
- [86] Y. M. Wazery, E. Saber, E. H. Houssein, A. A. Ali, and E. Amer, “An efficient slime mould algorithm combined with k-nearest neighbor for medical classification tasks,” *IEEE Access*, vol. 9, pp. 113666–113682, 2021.
- [87] Y. Zheng et al., “A novel hybrid algorithm for feature selection based on whale optimization algorithm,” *IEEE Access*, vol. 7, pp. 14908–14923, 2019.
- [88] M. Al-Zoubi, H. Faris, J. Alqatawna, and M. A. Hassonah, “Evolving support vector machines using whale optimization algorithm for spam profiles detection on online social networks in different lingual contexts,” *Knowledge-Based Syst.*, vol. 153, pp. 91–104, 2018.
- [89] H. Mohammadzadeh and F. S. Gharehchopogh, “A novel hybrid whale optimization algorithm with flower pollination algorithm for feature selection: Case study Email spam detection,” *Comput. Intell.*, vol. 37, no. 1, pp. 176–209, 2021.
- [90] J. Li, Y. Gao, K. Wang, and Y. Sun, “A dual opposition-based learning for differential evolution with protective mechanism for engineering optimization problems,” *Appl. Soft Comput.*, vol. 113, p. 107942, 2021.
- [91] Tan, X. Wang, and Y. Wang, “Dimensionality reduction in evolutionary algorithms-based feature selection for motor imagery brain-computer interface,” *Swarm Evol. Comput.*, vol. 52, no. April 2019, p. 100597, 2020.
- [92] F. Moslehi and A. Haeri, “A novel hybrid wrapper–filter approach based on genetic algorithm, particle swarm optimization for feature subset selection,” *J. Ambient Intell. Humaniz. Comput.*, vol. 11, no. 3, pp. 1105–1127, 2020.
- [93] S. Sahmoud and H. R. Topcuoglu, “A general framework based on dynamic multi-objective evolutionary algorithms for handling feature drifts on data streams,” *Futur. Gener. Comput. Syst.*, vol. 102, pp. 42–52, 2020.
- [94] Deniz, M. Angin, and P. Angin, “Evolutionary multiobjective feature selection for sentiment analysis,” *IEEE Access*, vol. 9, pp. 142982–142996, 2021.
- [95] W. Rathasamuth and K. Pasupa, “A modified binary flower pollination algorithm: A fast and effective combination of feature selection techniques for SNP classification,” in 2019 11th International Conference on Information Technology and Electrical Engineering (ICITEE), 2019, pp. 1–6.
- [96] C. Yan, J. Ma, H. Luo, G. Zhang, and J. Luo, “A novel feature selection method for high-dimensional biomedical data based on an improved binary clonal flower pollination algorithm,” *Hum. Hered.*, vol. 84, no. 1, pp. 34–46, 2019.
- [97] R. Hans and H. Kaur, “Hybrid biogeography-based optimization and genetic algorithm for feature selection in mammographic breast density classification,” *Int. J. Image Graph.*, p. 2140007, 2021.
- [98] Robnik-Šikonja and I. Kononenko, “Theoretical and empirical analysis of ReliefF and RReliefF,” *Mach. Learn.*, vol. 53, pp. 23–69, 2003.
- [99] Guyon and A. Elisseeff, “An introduction to variable and feature selection,” *J. Mach. Learn. Res.*, vol. 39, pp. 1157–1182, 2003.
- [100] Shishkin et al., “Efficient high-order interaction-aware feature selection based on conditional mutual information,” in International Conference on Neural Information Processing Systems, 2016, pp. 4644–4652.
- [101] X. V. Nguyen, J. Chan, S. Romano, and J. Bailey, “Effective global approaches for mutual information based feature selection,” *Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, pp. 512–521, 2014.
- [102] Gu, Z. Li, and J. Han, “Generalized fisher score for feature selection,” in Conference on Uncertainty in Artificial Intelligence, 2011, pp. 266–273.
- [103] Masaeli, Y. Yan, Y. Cui, G. Fung, and J. G. Dy, “Convex principal feature selection,” in Proceedings of the 2010 SIAM International Conference on Data Mining, pp. 619–628.
- [104] “VLDB ’75: Proceedings of the 1st international conference on very large data bases,” 1975.
- [105] S. Kemp, “Digital 2021: Global overview report,” *Datareportal*, 2021.
- [106] H. Li and P. C. Y. Sheu, A scalable association rule learning heuristic for large datasets, vol. 8, no. 1. Springer International Publishing, 2021.
- [107] L. Jiang, R. Li, W. Wu, C. Qian, and C. C. Loy, “Deepforensics-1.0: A large-scale dataset for real-world face forgery detection,” in Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2020, pp. 2886–2895.
- [108] Tagawa, “An approach to chance constrained problems based on huge data sets using weighted stratified sampling and adaptive differential evolution,” *Computers*, vol. 9, no. 2, 2020.
- [109] M. Elkano, J. A. Sanz, E. Barrenechea, H. Bustince, and M. Galar, “CFM-BD: A distributed rule induction algorithm for building compact fuzzy models in big data classification problems,” *IEEE Trans. Fuzzy Syst.*, vol. 28, no. 1, pp. 163–177, 2020.
- [110] Segatori, F. Marcelloni, and W. Pedrycz, “On distributed fuzzy decision trees for big data,” *IEEE Trans. Fuzzy Syst.*, vol. 26, no. 1, pp. 174–192, 2018.
- [111] T. Mouad, L. M. Driss, and K. Mustapha, “Big data traffic management in vehicular ad-hoc network,” *Int. J. Electr. Comput. Eng.*, vol. 11, no. 4, pp. 3483–3491, 2021.
- [112] Sadowski, Z. Galar, R. Walasek, G. Zimon, and P. Engelseth, Big data insight on global mobility during the Covid-19 pandemic lockdown, vol. 8, no. 1. Springer International Publishing, 2021.
- [113] F. Villanustre et al., “Modeling and tracking Covid-19 cases using Big Data analytics on HPCC system platform,” *J. Big Data*, vol. 8, no. 1, 2021.
- [114] Liu, Y. Fang, D. Zhu, N. Ma, J. Pan, and M. Q. H. Meng, “A Large-Scale Dataset for Benchmarking Elevator Button Segmentation and Character Recognition,” in IEEE International Conference on Robotics and Automation (ICRA), 2021, pp. 14018–14024.

Implementation of a Data Protection Model dubbed Harricent_RSECC

Frimpong Twum¹, Vincent Amankona^{2*}, Yaw Marfo Missah³, Ussiph Najim⁴, Michael Opoku⁵

Department of Computer Science, KNUST, Kumasi, Ghana^{1,3,4}

Department of Computer Science, CUCG, Sunyani, Ghana²

Department of Computer Science, UENR, Kumasi, Ghana⁵

Abstract—Every organization subsists on data, which is a quintessential resource. Quite a number of studies have been carried out relative to procedures that can be deployed to enhance data protection. However, available literature indicates most authors have focused on either encryption or encoding schemes to provide data security. The ability to integrate these techniques and leverage on their strengths to achieve a robust data protection is the pivot of this study. As a result, a data protection model, dubbed Harricent_RSECC has been designed and implemented to achieve the study's objective through the utilization of Elliptic Curve Cryptography (ECC) and Reed Solomon (RS) codes. The model consists of five components, namely: message identification, generator module, data encoding, data encryption and data signature. The result is the generation of the Reed Solomon codewords; cipher texts; and generated hash values which are utilized to detect and correct corrupt data; obfuscate data; and sign data respectively, during transmission or storage. The contribution of this paper is the ability to combine encoding and encryption schemes to enhance data protection to ensure confidentiality, authenticity, integrity, and non-repudiation.

Keywords—*Elliptic curve cryptography; encoding; encryption; Reed Solomon; security*

I. INTRODUCTION

A. Background of Study

The advent of computerized systems and networks has been beneficial to organizations and has subsequently enhanced their operations. This has resulted in the generation of larger quantum of data to augment the activities of these organizations. Data produced by these organizations are considered a major resource, therefore resulting in organizations adopting strategies that can protect this important resource from being misused. As a quintessential resource, comprehensive techniques are provided and instituted by these organizations frequently to offer protection to this data [1].

Though protective mechanisms are instituted, there is also an increase in threats to undermine organizations' operations. Notwithstanding, the insurgence of adversary's attacks have consistently hampered the functional activities of organizations over the past years and was colossal during the COVID-19 pandemic era. To ameliorate this insurgence, organizations started scrambling for solutions to protect their data. In this regard, researchers began to also seek for

solutions to mitigate this insurgence of threats and attacks through the development of robust techniques and methods to prevent loss of data and unauthorized access and modification. It is on the basis of the aforementioned, that it is always important for industry and researchers to stay a step ahead of attackers in the preservation of data in transit and storage, hence, the call for this research.

As several research have postulated, most security systems have focused on either using encoding or encryption strategies/techniques to guarantee the safety and accuracy of data [2]. Whereas the encoding schemes add extra bits to the original data to aid in error detection and correction, in order to maintain the integrity of messages. The encryption schemes, on the other hand, ensure messages transmitted or stored are obfuscated to prevent unauthorized access and modification in order to maintain the authenticity and confidentiality of messages.

Examples of data encoding and encryption schemes include Reed Solomon codes, Reed-Muller codes, checksum, and AES, RSA, ECC, Blowfish among others have emerged to offer protection to data [3]. The encoding and encryption schemes are aimed at preserving the confidentiality, authenticity, integrity and non-repudiation (CAIN) of data. The utilization of encoding scheme over encryption algorithms, even though, offers security but it is ineffective to provide optimal protection due to unauthorized access or alteration of data as a consequence of adversaries' activities. This illicit access and modification of the data by the attacker causes data compromise. Also, the utilization of encryption algorithms offer protection to data but it is inadequate as a result of inadvertent modification, loss of data or hardware failures. Data can be destroyed by hardware failures, untrusted communication channels or attackers when accessed. This also leads to data compromise and therefore requires the application of different security measures to offer optimal security.

This research therefore focuses on how to harness the strengths of encoding and encryption security techniques to achieve a robust data protection system.

B. Aim of Study

The study's aim is to implement a data protection model by integrating Elliptic Curve Cryptography (ECC) and Reed Solomon (RS) coding schemes.

*Corresponding Author.

To achieve the stated aim the following questions are raised:

- 1) Can ECC and RS codes be used to ensure secure data transmission and secure data storage?
- 2) Can a proposed Harricent_RSECC data protection model enhance data security by ensuring an uncompromised data transmission and data storage?
- 3) Will the implementation of the proposed Harricent_RSECC data protection model offers the required security to data?

II. ELLIPTIC CURVE CRYPTOGRAPHY (ECC)

Elliptic Curve Cryptography (ECC) is a contemporary group of public-key cryptosystems premised on the algebraic structures of elliptic curves over finite fields and the complexity of the Elliptic Curve Discrete Logarithm Problem (ECDLP).

An ECC curve can be illustrated as a curve that intersects two lines on a graph. This type of curve is determined by the properties of the mathematical group consisting of set of values for which operation on two of its members produces a third member [4] depicted in Fig. 1. Multiplying a point by a number on the curve produces an additional point on the curve, but finding what number has been used is very difficult, although parties involved know the original point and results. ECC uses elliptic curves in which elements of a finite field are all limited to variables and coefficients. The ECC is mathematically represented using the Weierstrass form of an elliptic curve denoted in (1) as follows $y^2 = x^3 + ax + b$ where $4a^3 + 27b^2 \neq 0$. Each "a" and "b" value has an elliptic curve that is different.

$$y^2 = x^3 + ax + b \quad (1)$$

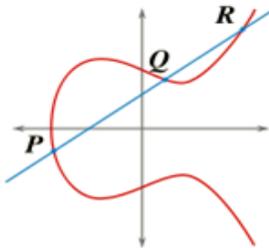


Fig. 1. ECC Representation.

ECC constructs all of the substantial functionalities of asymmetric cryptosystem, including encryption, signatures, and key exchange. ECC cryptography is regarded as an effective modern replacement to the RSA cryptosystem since it utilizes smaller keys and signatures than RSA for the same level of security and offers extremely fast key generation, key agreement, and signatures.

A. ECC Keys, Algorithms, Curves and Key Length

In the ECC, the composition of private keys are generally numbers within a range of integers (usually 256-bit integers), thereby making it easier and faster for private key generation. The public key on the other hand is generated from points which lay on an elliptic curve, usually a pair of integer

coordinates (x, y). Ultimately, a shared key is a public key that is derived after multiplying the private key of the sender to the public key of the receiver and vice versa [5].

Besides, basing on the mathematical properties of elliptic curves over finite fields, the elliptical cryptography offers varied sets of algorithms. Three major categories of ECC algorithms are available consisting: signature algorithms such as elliptic curve digital signature algorithms, fast elliptic curve digital signature algorithms and Edwards digital signature algorithms; encryption algorithms for instance elliptic curve integrated encryption scheme and ElGamal Encryption using ECC (EECC); and lastly is the key agreement algorithms including elliptic curve Diffie Hellman X25519 and Fully Hashed Menezes-Qu-Vanstone (FHMV) [6].

Moreover, diverse sets of curves exist that elliptic curve crypto algorithms can utilize to achieve different purposes. Goals such as determining the level of cryptographic strength (security), the length of key and the performance are rational for implementing a different set of curves. Consequently, every curve consists of the following parameters (curve name; size of the field/key; the cryptographic strength – expressed as ratio of field size to 2; and the speed – also expressed as ratio of operations to seconds). Examples of ECC curves and their corresponding key sizes include “curve secp192r1 with 192-bit, curves secp256k1 and Curve25519 with 256-bit, curve P-521 with 521-bit among others” [6].

Therefore, it can be advanced that digital signature, encryption and key agreement algorithms utilizes any of the curves above for computations which heavily rely on the difficulty of ECDLP to provide adequate level of security to data while performing with minimal key length.

B. Key Benefits of Elliptic Curve Cryptography

The benefits of using elliptic curves stems from the fact that with shorter keys, equivalent levels of protection can be obtained compared with other algorithms. The ECC cryptosystem is useful in our contemporaries with massive upsurge in the development and usage of mobile devices. Thus, as the use of smartphones grows, more robust encryption is necessary for businesses to meet the increasing security requirements [7]. Some of the benefits are discussed below:

1) *Stronger keys:* ECC represents the latest encryption method, providing more security in elliptical curve cryptography. The underlying math problem of the ECC algorithm implies it is more difficult for hackers to crack compared to RSA and DSA, which makes the ECC algorithm more reliable than conventional methods for websites and infrastructure [8]. ECC algorithms rely on the difficulty of ECDLP to provide adequate level of security to data while utilizing minimal key length [5]. This invariably has improved performance and enhanced storage requirements.

2) *Short key size:* Comparably, Because ECC keys are substantially smaller than DSA/RSA keys; the size of the public and private keys in ECC is relatively short. As a result, the National Institute of Science and Technology (NIST) recommended that ECC can employ substantially lower

parameters for the same degree of security bits as RSA/DSA [9]. The choice for these algorithms is that elliptic cryptography uses lesser keys to achieve same security levels contrary to the other PKCs. For instance, the RSA/DSA technique requires a key length of 7680 bits to achieve 192 bits of security, whereas ECC requires a key size of 384 bits. Furthermore, the RSA/DSA algorithm requires a key size of 15360 bits to obtain 256 bits of data protection, whereas ECC requires a key length of 512 bits.

As a result, processing times are reduced while memory and bandwidth requirements are limited. ECC is especially suitable for applications with limited memory, bandwidth, and/or computing capability, and its use is expected to increase in this area.

III. REED SOLOMON (RS) CODING SCHEME

A. The Theory of Reed Solomon (RS) Coding

The RS code is a systematic, linear cyclic and non-binary block code. During RS coding, redundant symbols are created and appended to the symbols of the message by using a polynomial generator [10]. The position and magnitude of the error in the decoder are determined using the same polynomial generator [11]. The correction is then applied to the code received. RS coding is commonly used for error detection in a variety of communications and computer infrastructures, including storage, wireless or mobile communications, satellite communications, digital television, and high-speed modems [12].

RS codes were opted over other error detection and correction codes because of their faster decoding capabilities, i.e., their ability to detect and/or correct significant numbers of omitted or compromised data items; and the fact that they require the fewest additional error correcting codes bits for a known number of data bits [10]. Fig. 2 provides a framework representation of RS encoding/decoding process:

Reed-Solomon codes are by far the most extensively utilized for burst error correction [13]. The benefit of utilizing RS codes is that the likelihood of an error persisting in the decoded data is (generally) substantially lesser than if RS codes are not employed. Coding gain is a common term for this benefit. Because Reed Solomon code has a high coding rate, it is suited for a broad array of applications, comprising storage and transmission [12].

B. Properties of RS Code

RS code is specified as $RS(n, k)$ with s -bit symbols [10]. Given a symbol size s , the maximum codeword length (n) for a Reed-Solomon code is $n = 2^s - 1$. For instance, an 8-bit symbol will produce a maximum length of 255 bytes. This signifies that the encoder creates a n symbol codeword by adding parity symbols to k data symbols of s bits each. There is a total of $n - k$ parity symbols, each with s bits. A Reed-Solomon decoder can fix up to t symbols in a codeword that have errors, where $2t = n - k$. A typical Reed-Solomon codeword (known as a Systematic code because the data is not altered and the parity symbols are attached) is shown in Fig. 3 [14].

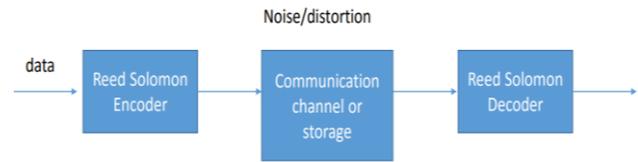


Fig. 2. Framework of RS Scheme.

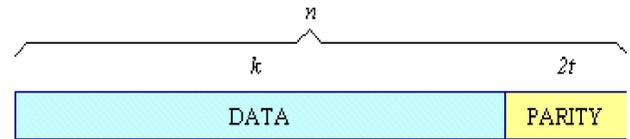


Fig. 3. Reed Solomon Codeword Generation.

A code such as Fig. 3 can detect and correct up to $(n - k)/2$ or t symbol where each symbol represents an element within the finite fields $GF(2^m)$. This implies that any t symbols that may be corrupted can still be recovered from the original message [15].

An $RS(255, 223)$ with 8-bit symbols is an example. Each codeword is made up of 255 bytes, with 223 bytes of data and 32 bytes of parity. The following can be derived from the code:

$$n = 255; k = 223; s = 8; 2t = 32, t = 16$$

As a result, the decoder can automatically correct any 16 symbol errors in the codeword: that is, errors of up to 16 bytes can occur anywhere in the codeword.

In Fig. 4, the sender uses the RS encoder to encode the message in a codeword and transmits it through the communication channel. Channel noise and other disorders may disrupt the codeword and corrupt it. This corrupted codeword comes to the recipient end (decoder) and transfers the tested message to the receiver. If the error caused by the channel is larger than the decoder's error correction capability it may result in a decode failure. Decoding errors occur when a codeword has not been passed and a decoding error leads to an incorrect message [16].

The representation of an instance of an RS protected communication channel while transferring data is illustrated in Fig. 4.

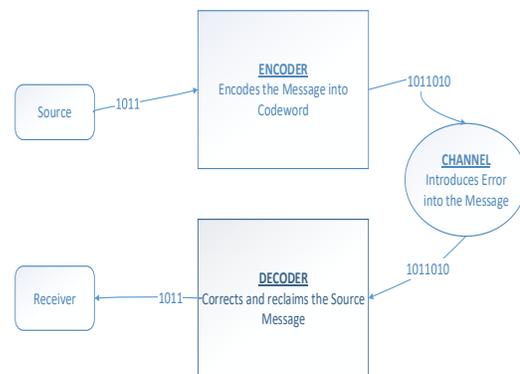


Fig. 4. Reed Solomon Protected Channel.

IV. REVIEW OF RELATED WORK

This section explores the extant works of ECC and RS. Elliptic curve cryptography (ECC) is by far the most effective public-key option for providing security services to devices with limited resources and it has been employed in a variety of business applications [5]. ECC, since its emergence has been considered the preferred option with notable efficiency for ensuring authenticity, encryption, signatures and key agreements [5]. To achieve these, He et al. deployed an ECC authentication model in a smart grid environment to obfuscate smart meter anonymity [17]. However, Sadhukhan et al pointed some security flaws in [17] such as internal and masquerading attacks [9]. Besides, by utilizing ECC and image steganography to maintain the legality and accuracy of medical health data, Eshraq et al.'s model obfuscated health data from being accessed by unauthorized access [18]. To ensure ECC's applicability in diverse contexts, the authors [19], [20], [21], [22] and [23] utilized ECC to achieve confidentiality, integrity, non-repudiation and authenticity levels of security.

On the contrary, the establishment of a reliable communication channel is very necessary as there exist security breaches within satellite or telephone channel which can compromise data security [10]. To overcome these security breaches many techniques of correcting errors have been introduced over time. One of such method is RS code which is a key non-binary BCH coding sub-class. These are useful cyclic codes utilized in detecting and correcting burst errors. RS codes have been prevalent due to its simplicity in encoding and well-structured decoding capabilities. In appraising the efficiency of RS codes, Wonshik & Jae-Yeon appraised the performance of RS(255, 239) in a smart transmission medium and an intelligent system [12]. The implementation of their system depicted RS code efficiency over a chaotic communication channel as the higher the codeword length, the greater the bit error rate improvements. Besides, Mounika [10] postulated that there exist security breaches on a satellite or telephone channel. Therefore, the establishment of a reliable communication channel is very necessary. To this end, [10] implemented a modified version RS codes for performing error corrections. Further, studies by [13] indicated that the decoding capabilities of RS codes have been efficient against deletion errors. And it was validated through the simulation results by [24] where the proposed RS decoder achieved a higher coding gain in contrast to algebraic decoding methods.

The identifiable gap is the ability to combine both techniques to achieve a robust security model and that's the goal of this study.

V. IMPLEMENTATION OF THE HARRICENT_RSECC MODEL

A. Conceptual Framework of an Efficient Data Security

Quite a number of studies have been carried out relative to mechanisms that can be deployed to enhance data protection. However, most of these studies have either focused on encryption or encoding schemes as postulated in [2]. The ability to integrate these two techniques and leverage on their strengths to achieve adequate data protection has been the

major concern in this study. To this end, the Harricent_RSECC data protection model has been designed and implemented to achieve the objective of this study through the utilization of Elliptic Curve Cryptography (ECC) and Reed Solomon (RS) codes. The prime goal for integrating ECC and RS is to achieve a fast, small, and portable cryptographic protocol, which would support elliptic curve digital signature generation and verification together with data reconstruction.

Fig. 5 shows the design of the Harricent_RSECC data protection paradigm, which integrates RS encoding and two variants of ECC to serve a purpose of improving data protection while achieving data integrity and confidentiality. Generally, the implementation of the model is achieved by the following steps.

B. Metrics used for Harricent_RSECC Data Protection Model

As a defensive system, it is always important to identify any messages or data received. Since messages come in different forms, understanding the different message types received play a crucial role when identifying the message/data received. The process for an unwanted message to be detected or identified is important to ensure the right defensive mechanisms are applied.

Having identified the type of message, a second defense module, the generator module, ensures proper defense by identifying which encoding level is crucial in identifying and correcting the right amount of errors that may occur in the system. This module also provides the chunking of data to ensure faster processing of the messages received.

Thereafter, messages received are transferred for security mechanisms to be applied to the message. The mechanisms are to encode, encrypt and sign the messages. These three effective security mechanisms will prevent unauthorized access to the message received.

As different attacks can be performed, each stage in the framework provides a defensive approach in protecting the message received.

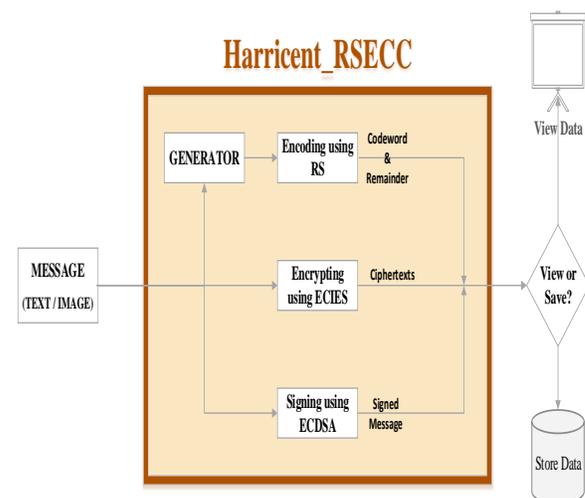


Fig. 5. Schematic Diagram of Harricent_RSECC Data Protection Model.

Therefore, the metrics that defined the functional requirements for Harricent_RSECC are:

- 1) The model must be able to identify the message type in order to perform the proper analysis. This maintains the scope of the study (text and images) prevent unclassified messages from being uploaded.
- 2) The model must be able to identify the message size/length in order to apply appropriate RS coder levels.
- 3) The model should be able to detect if a message has been compromised and correct eventually.
- 4) The model should be able to obfuscate the message and prevent unsolicited access and modification.
- 5) The model should be able to maintain the integrity of the message and associate messages transmitted or stored with the sender (i.e., ensuring non-repudiation).

The primary focus of this model is the ability for the system to identify a message type received, chunking of large messages for faster processing, obfuscation of the message, signing of the message, detection and correction of compromised messages. This work attempts to provide a protective mechanism which organizations can rely on to protect their resources.

C. Implementation of Harricent RSECC Components

From the design of the model, the following principal components are implemented: Message Identification, Generator Module, Encoding, Encryption, and Signature.

1) *Message identification:* A message is submitted by the user into the Harricent_RSECC model. This phase is to identify the type of message that is permissible to be uploaded into the model, be it a text or an image. This module consists of a rule-based component where rules are defined based on the list of file types and type of file uploaded or sent by the user.

For example, a typical rule might be `if(filetype==filetypeslist)` return true. The module identifies the message/data based on the type of file received/uploaded. The received message/data is processed and categorized to compute a value M_t based on two outputs, text (Dt) and Image (Di). Messages are identified based on $f(M_t)$ as follows:

$$f(M_t) = \begin{cases} D_t, & \text{if}(M_t = \text{filetypeslist}_t \wedge \text{filetype} = \text{filetypeslist}), \\ D_i, & \text{if}(M_t = \text{filetypeslist}_i \wedge \text{filetype} = \text{filetypeslist}). \end{cases}$$

If new message types are identified, there is the need for it to be added to the file type list to improve the efficiency of the system. If new message types are identified, there is the need for it to be added to the file type list to improve the efficiency of the system.

Files uploaded are checked by a function to prevent unclassified files from being uploaded.

```
#function file upload
```

```
if (filetype = ".txt" or filetype = ".doc" or filetype = ".docx" or  
filetype = ".png" or filetype = ".jpg" or filetype = ".jpeg" or  
filetype = ".gif");
```

```
upload file
```

```
else
```

```
print("file type not supported")
```

2) *The Generator Module (GM):* This module prepares image and text files for the RS encoding scheme to be applied. It determines the length of message, chunks message based on encoder level and generates ASCII codes from message. The received message could be preprocessed by this module to determine the message's length. Message length, M_L , determination effectively enables how the message can be encoded. If a message, M , is presented as a number of characters w_1, w_2, \dots, w_n , then M_L is determined by w_n , where n is the last returning value, resulting in $M_L = n$. This module also takes into accounts the coder level during processing to determine the bits-type needed during encoding. The size of the bits-type is determined based on the message length M_L . The bits size, k , of a message is the bits-type to be used during RS encoding. Selection of the bits-type for a message is determined based on

$$f(M_k) = \begin{cases} 4bits, & \text{if}(M_L < \alpha), \\ 8bits, & \text{if}(M_L \geq \alpha) \end{cases}$$

where α is a pre-defined threshold value of thirteen characters.

The algorithm for determining the length of text messaging is presented below.

```
#def determine the length of message.
```

```
Take the length of message using the length library.
```

```
#def set bit size.
```

```
if messagelength < 13 and Bit_size=="8-bits":
```

```
Display a message that coder level must be either 'RS(15, 11)  
and RS(15, 9).
```

```
Set bit_size to 4-bits.
```

```
Else.
```

```
Pass
```

After the message length has been determined, a decision has to be made where the received message will be chunked (1) or not-chunked (0), utilizing the encoder level. The encoder level utilizes message length CML and codeword length CWL in its operation. All encoder levels are in the format: $CL(CW_L, CM_L)$, where CL is for coder level. Chunked messages M_c are determined based on.

$$f(M_c) = \begin{cases} 0, & \text{if}(M_L \leq CM_L), \\ 1, & \text{if}(M_L > CM_L) \end{cases}$$

The algorithm for chunking a message is presented below.

```
#code for chunking data.
```

```
#def chunkmydata(self, string, n):
```

```
chunks = [string[i:i + n] for i in range(0, len(string), n)].
```

```
return chunks.
```

The last operation of this module is the conversion of a message to its ASCII format. Each character (C) in the message is converted to its *ASCII code* (AC) (ie. $C_1 \Rightarrow AC_1, \dots, C_n \Rightarrow AC_n$).

The algorithm for messaging to their equivalent ASCII is presented below.

```
#def convert message length to ASCII codes.
```

```
letters = take the ordinal of the message characters to give  
their ASCII Codes.
```

3) *Data encoding*: This module utilizes the RS error correcting code to encode messages prepared from the GM to generate codeword. This is to guarantee that messages in transit or storage are devoid of corruption or contain errors by appending redundant data to it. The framework employed multiple encoder levels to encode. The RS code is built on finite fields, which have the feature that any computation on field elements always returns an element in the field-set [14].

To generate an RS codeword, a polynomial generator $g(x)$ is used with the following notations: information block $i(x)$, codeword $c(x)$ and a primitive element (α) of the field. A polynomial generator is illustrated in (2) below:

$$g(x) = (x - \alpha^i)(x - \alpha^{i+1}) \dots (x - \alpha^{i+2t}) \quad (2)$$

with the codeword generated as follows (3):

$$c(x) = g(x)i(x) \quad (3)$$

Following steps are employed in the generation of codewords and remainders:

a) An encoder level (RS) must be selected. To encode, Reed Solomon coder makes use of different encoder levels. Eight (8) RSCoder levels (RS) were selected ranging from (15, 9) to (255, 251) as follows:

```
RSCoder(15, 9), RSCoder(15, 11), RSCoder(53, 37),  
RSCoder(255, 223), RSCoder(255, 239), RSCoder(255, 251),  
RSCoder(255, 247), RSCoder(255, 191).
```

The above RS encoder levels determine the message (RS_M) and codeword (RS_C) lengths.

b) Message (M) length is determined based on the encoder level (RS) selected.

```
Message length = len(Message)
```

c) If message (M) length is less than or equal to the message (RS_M) length of the encoder, message (M) is encoded and a remainder (R) is generated.

d) A message (M) length that exceeds the encoder level message length (RS_M), message (M) is chunked based on the

message (RS_M) length of the encoder level using a chunk function explained above. Each chunked message is encoded and a remainder (R) generated for each.

```
Chunked Data = (M,  $RS_M$ ).
```

```
Chunked Data Length = len(Chunked Data).
```

```
If Message length >  $RS_M$ :
```

```
Chunked Data = (M,  $RS_M$ ).
```

```
codeword = rs4a.encode(Message).
```

```
remainder = codeword[-6:].
```

```
remainderhex = remainder.encode().hex().
```

4) *Data encryption*: The Encryption module provides obfuscation and protection of message/data. The framework used the ECIES to accomplish this. In an unsecured channel, ECIES will safeguard the contents from being read or altered with by unauthorized parties. During implementation, the ECIES combines the Secp256k1 curve with the Advanced Encryption Standard – 256 – Galois Counter Mode (AES-256-GCM) to offer the needed security. The AES-256-GCM algorithm is a symmetric encryption algorithm with a 256-bit key size [25]. To achieve ECIES encryption the following steps are followed:

a) *Generation of Key Pairs*: The first step in ECIES is to generate private (PRk) and public key (PUk) pairs. In the elliptic curve, private keys (PRk) that are generated are in integers and public keys (PUk) are points on the curve. Therefore, to encrypt file contents, the Harricent_RSECC model generated both public (PUk) and private keys (PRk). A private key (PRk) was created by obtaining 32 bytes from randomly generated numbers (R). A new private key (PRk) is generated if there is none; otherwise the private key already generated will be loaded. A private key (PRk) is just a sequence of raw bytes. To ensure better security, small numbers must not be chosen as private keys.

```
PRk = rand()
```

```
If PRk = NONE then
```

```
PRk = rand()
```

```
Else
```

```
PRk
```

Example: the limit of the integer number (N) as agreed by the communicating parties.

```
N =
```

```
10015792089237316195423125709850085568790718528137  
56427907490438260516314
```

```
PRk = random.randint(1, N)
```

```
PRk =
```

```
95193991353039061015378100235213857196721754949810  
9375643560636694135579
```

The public key, on the other hand, is determined by multiplying a point on the elliptic curve (G) and the private key (PRk).

$$PUk = PRk \times G$$

Therefore,

$$PUk = (PRk \times Gx) + (PRk \times Gy)$$

b) Generation of Shared Secret Keys: The second step of using the ECIES was to create a shared secret key by multiplying the private (PRk) and public (PUk) key using Elliptic Curve Diffie-Hellman (ECDH) algorithm. The ECDH algorithm is principally used to prevent eavesdropping by facilitating the exchange of the AES shared secret key. This shared key is a public key derived after multiplying the private key of the sender to the public key of the receiver and vice versa. Thus,

$$\text{Shared Secret key (S)} = PRk \times PU$$

c) Key Derivation Function (KDF) - AES Key and HMAC Key: The third step of the ECIES algorithm utilized the shared secret key to derive an Advanced Encryption Standard (AES) key and a Hash-Based Message Authentication Code (HMAC) key via a Key Derivation Function (KDF). The KDF component provides another layer of security to prevent man-in-the-middle attack which is a major issue in ECDH. The KDF ensures the hashing of the shared secret key (S) generated using SHA256 algorithm. The steps involved in implementing the key derivation function (KDF) are:

- i) Convert Shared key (S) to bytes
- ii) Using the SHA256 hash library, the bytes are hashed to produce a hash key (H).

The AES key derivation is based on the key length and the hash (shared) key (H). AES-256-GCM has the key length of 256 bits. To generate an AES key that was used for encryption, the hash (shared) key (H) length should be equal to 256 key bits length.

$$\text{AES Key Length} = 256 \text{ bits}$$

$$\text{AES Key} = H [: \text{AES Key Length}]$$

// AES key picks the first part of shared //key while the latter is for the HMAC key

More so, HMAC secret key utilized was also based on the key length and the hash (shared) key (H). HMAC's cryptographic strength depends on the size of its output as well as the size and quality of the key. HMAC-SHA256 operates on a bit size of 256. This size produces a sizable output and key. To generate HMAC secret key:

$$\text{HMAC Key Length} = 256 \text{ bits}$$

HMAC Secret Key = H [HMAC Key Length:] // HMAC key picks the later //part of the shared key

d) Message Encryption: The fourth step in adopting the ECIES is message encryption. To encrypt a message, a random initialization vector (IV) was created and XOR'd with the

message in addition to the AES key to generate a ciphertext. Initialization vector (IV) is an abbreviation of "nonce" which connotes the number used once. The randomly generated bits were based on the AES block size and the number of bits in the key length.

The generated IV is converted to bytes depending on the AES block size.

$IV = \text{random.getrandbits}(\text{AES block size} * \text{number of bits per byte})$.

$IV = IV.to_bytes(\text{AES block size})$.

$\text{Ciphertext (C)} = (IV \text{ XOR message, AES Key})$.

e) Hash-Based Message Authentication Code (HMAC): In the fifth step, HMAC was created to be used for initialization vector (IV) and Ciphertext (C). To create HMAC, two keys are generated from the HMAC Secret Key. The two keys are named, inner key (Ik) and Outer key (Ok). The first 256 hash (Fh) value is generated from the message and Ik. A second 256 hash value is generated from Fh and Ok. These keys are sent to the recipient for verification/decryption.

f) Transmitting PUK, IV, C and HMAC: The final step involves the sending of PUK, Initialization Vector (IV), Ciphertext (C) and Hash-Based Message Authentication Code (HMAC). The transmission is through a communication channel to the server facility for storage.

Encrypted data or ciphertext sent (C_s) is decrypted using the private key PR_k ,

$$\text{Message}_{\text{decrypted}} = (C_s, PR_k)$$

5) Data signature: This component uses Fast ECDSA to digitally sign files to uphold the integrity of the message. In comparison to other types of digital signatures, the Fast ECDSA technique is designed to conduct fast elliptic curve cryptography. In comparison to ECDSA, it takes very little time to perform its instructions. In the elliptic curve, private keys are integers, and public keys are points on the curve. To offer 128 bits of security, the curve adopted is P256/secp256r1. Creation curves are done using Weierstrass form: $y^2 = x^3 + ax + b(\text{mod } p)$.

The Fast ECDSA algorithm employed to digitally sign messages combines P256 / secp256r1 curve and SHA3_256. The curve is imported for use in this study as shown in the example below:

Example:

curve = ec.SECP256R1()

The following are the steps involved in the implementation of Fast ECDSA algorithm:

a) Generation of Key Pairs: Private (EPRk) and Public (EPUk) keys were generated to sign and verify messages by using P256 curve. Private (EPRk) keys are integers that were randomly generated and Public (EPUk) key were derived by multiplying the EPRk and a point on the curve (GE). Public

VI. DISCUSSION

The Harricent_RSECC data protection model provides data owners with the guarantee of data security in transit and storage while preserving data validity. It contributes to the field of security by incorporating encryption, encoding, and signature techniques to provide confidentiality, authenticity, integrity, and non-repudiation levels of security. Should an adversary intercept any of the ciphertexts, the model ensures data secrecy by utilizing ECIES to prevent unauthorized access and eavesdropping. The ECC key size for the Harricent_RSECC data protection model was 256. This provides a 128-bit security equivalent (RSA/DSA will use a key length of 3072). Furthermore, the model employs Fast ECDSA to digitally sign a message, ensuring its integrity and non-repudiation. Fast ECDSA generates hash values from the message which are used for signing the messages and verifying the authenticity of the ciphertexts. The model additionally leverages RS codes to identify and correct errors that arise in order to provide the capability of retrieving corrupted data files that occurred during transmission or storage, which further enhances data protection in transit or at storage.

The model affirms the studies conducted by [19], [20], [25], [26], [27], [21] that applied elliptic cryptographic algorithms to encrypt and decrypt the selected data and this guarantee the safety of private information, sensitive data, and can enhance the security of communication between computer systems. The Harricent_RSECC model through the use of Fast ECDSA validates the identity of the sender that transmitted the message which upholds the studies conducted by [19], [20] [28]. By employing Fast ECDSA, the content of the message cannot be altered without detection. This ensures the integrity of the message is maintained and, moreover, the signer cannot deny association with the signed content. Moreover, studies conducted by [13], [12], [10], [23] through the utilization of RS codes endorses the benefits of error detection and correction that may emerge.

A comparative discussion supported with existing literature is presented in this section. The security of the Harricent_RSECC data protection model was evaluated using five metrics: Message Identification, faster processing through chunking, obfuscation, detection and correction of compromised messages and signature of message. This is to ensure that the confidentiality, authenticity, integrity and non-repudiation (CAIN) levels of security are provided to the data. By obfuscating message, the confidentiality and authenticity of information stored on computer systems or transferred between its users across the internet is assured. By encoding and decoding the message, the integrity of the message is maintained through the addition of redundant data. By signing the message, the integrity and non-repudiation of the message is ensured. The selected metrics, as adduced in chapter 3, aids in appraising the security potentials of the system.

In analyzing the studies of [19] and [20], it was observed that their system partly used the message identification metric by concentrating on either text or image on separate studies. To ensure fast computational process, their studies utilized message groupings to chunk messages. However, their

adopted message groupings will slow the processing of messages as compared to the chunking component (generator module) in the Harricent_RSECC model. In our study, the GM chunks each message based on the RS coder level which leads to faster encoding processing. Even though, the authors utilized the ECC to obfuscate the messages, the two distinct studies lack the functionality of detecting and correcting of compromised messages in an event of error occurrence. More so, [19] and [20] studies failed to offer and maintain integrity of messages by not signing and associating messages transmitted or stored to the sender (i.e., ensuring non-repudiation); which have been ensured through fast ECDSA in Harricent_RSECC model.

In a related study, [23] developed a technology that allows messages to be hidden based on numerical ruler-bundle. It's worthy to note that message hiding is one of the reliable methods of preserving and ensuring the CAIN. Notably, [23]'s study partially made use of only one part of message identification (i.e., image) in comparison with Harricent_RSECC model. Analysis of [23]'s technology also reveals that the authors failed to decompose the information extracted from the image as their system read each stream of input sequence to determine its encoding alphabet; however, their technology hampers efficient computational processes. Their studies enable the hiding of an image inside another image in addition to noise tolerant codes. The model of noise-tolerant codes provided an opportunity to correct up to 25% of errors in the code word. Though, the technology has the capacity to detect and correct errors in comparison with Harricent_RSECC Model, but it lacks the capacity to ensure non-repudiation as their technology does not sign messages that are transmitted and stored.

Also, a study conducted by [21] concentrated on text messages the author's message identification at the exclusion of images. Besides, [21]'s study failed to highlight the imports of chunking which Harricent_RSECC model dwells because chunking is beneficial to ensuring faster computational processes. Like the Harricent_RSECC model, [21]'s model obfuscates the content of the message through the utilization of ECC. However, it lacks the ability to sign, detect and correct compromised messages; hence integrity and non-repudiation levels of security are not maintained.

More so, [22] designed a secure model for data protection in the cloud but the system lacks the ability to differentiate the types of data the study seeks to secure. Data identification is consequential in ensuring CAIN because each data has its own way of processing to offer the needed protection. Also, [22]'s model was not clear on how message chunking is being handled; however, much concentration was given to usability in the study. While the latter is good, the former is equally important to ensure fast computation. More so, [22]'s model obfuscated messages, but the methods employed are AES and RSA, which in literature this study has adduced that the benefits ECC far outweighs RSA. Again, the [22]'s study indicated the concepts of non-repudiation and integrity but the methods in achieving those security metrics are not explicit. Lastly, their system lacks the ability to detect and correct compromised messages which inadvertently affect the integrity of the transmitted or stored data.

TABLE I. COMPARATIVE ANALYSIS OF SOME EXISTING WORK WITH HARRICENT_RSECC MODEL

Authors	MI		C	O	EDC	S
[19]		Image	Yes	Yes – ECC	No	No
[20]	Text		Yes	Yes - ECC	No	No
[21]	Text		No	Yes, ECC	Not	No
[22]	No message identification		Not clear	Yes, AES and RSA	No	Non repudiation and integrity are mentioned but method not clear
[23]		Image	No	Yes - numeric ruler bundle	Yes - numeric ruler bundle	No
Harricent_RSECC	Texts and Images		Yes	Yes - ECIES	Yes - RS	Yes – Fast ECDSA

Table I presents a comparative analysis of some existing work with Harricent_RSECC model in achieving secured system by basing on the stated metrics:

MI – Message Identification,

C – Chunking,

O – Obfuscation,

EDC – Error Detection and Correction,

S – Signature,

From the levels of security espoused, it can be asserted that it is only the Harricent_RSECC model that combines the functionalities of encrypting, encoding and signing to achieve confidentiality, authenticity, integrity, non-repudiation. Therefore, the Harricent_RSECC data protection framework offers the data owner the assurances of data protection in transit and storage, while maintaining the validity, integrity and confidentiality of the data. It is valid, therefore, to assert that Harricent_RSECC model is the efficient framework comparatively and offers protection to text and image datasets by using both ECC and RS encoding protocols.

To advance this section, the questions posited in section 1.3 are discussed.

A. Can RS and ECC be used to Ensure Secure Data Transmission and Secure Data Storage?

The characteristics of RS and ECC are advanced in Sections 2 and 3. To offer protection to data, several data encoding and encryption techniques have been proposed. In the case of the encryption, such schemes are utilized to obfuscate messages to prevent unauthorized access and modification to protect the validity and secrecy of messages transmitted/stored. While the encoding scheme adds extra bits to the original data to aid in error detection and correction in order to keep the messages from corruption or damage.

This study evaluates the RS encoding scheme and ECC encryption technique to create a data protection model. Data in transmission and/or storage are susceptible to attacks or security threats. Most security systems focus on either extra bit of data to messages or obfuscate the message. The utilization of one scheme over the other offers security but it is ineffective to provide optimal protection due to the inadvertent

loss of data as a consequence of adversary’s activities or hardware failures. Therefore, it is essential to combine both encoding and encryption strategies for ensuring effective data protection.

As a public key cryptography, ECC characteristically consists of a public key and a private key which augment the communication between the parties involved. As a result, ECC use significantly smaller parameters than RSA/DSA to achieve the same level of security.

B. Can a Proposed Harricent_RSECC Data Protection Model Enhance Data Security by Ensuring an Uncompromised Data Transmission and Data Storage?

The Harricent_RSECC model and implementation (shown in Fig. 8) achieved the study’s purpose of providing enhanced data confidentiality and integrity model by using Reed-Solomon encoding scheme and elliptic curve cryptography. This ensures that the data owner participates in the provision of security to the datasets before storing or outsourcing it to a storage infrastructure [21].

Firstly, from the implementation of the model, a message must be encrypted Using ECIES to inhibit unauthorized access and intrusions if any of the ciphertexts are intercepted. Secondly, prior to transmission, the message is signed by computing a hash value to enable the receiver to authenticate the validity and genuineness of the source message. This aids in detecting any compromises that ensued during transmission by untrusted channel or by an attacker.

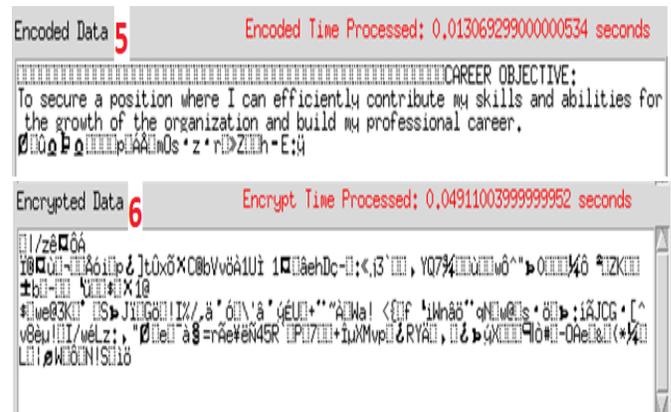


Fig. 8. A Snapshot of an Encoded and Encrypted Data.

Moreover, in the event of data compromise, the Reed Solomon scheme is utilized to decode and correct the corrupted message. Thus, at the receiving end, the receiver validates the validity and accuracy of the message by computing the hash value on the decrypted message to verify the authenticity of the message. The unmatched values computed from the verification algorithm would render the invalidity of the message.

C. Will the Implementation of the Hybrid Data Protection Scheme Offer Necessary Security to Text and Image Datasets?

The Harricent_RSECC data protection model is a robust model that offers higher security levels to texts and image datasets as it combines the strengths and efficiency of two outstanding data protection schemes. The implementation of the Harricent_RSECC data protection model offers necessary security to both image and text dataset including confidentiality, integrity and non-repudiation. From Appendix I, the Harricent_RSECC model provides confidentiality of the owner's data at the first level of security through encryption by using the ECIES. Again, at the second level of security, the ciphertexts are signed using the fast ECDSA to sign, authenticate and validate the source of the transmitted data. The Harricent_RSECC model provides protection from data loss or corruption using the Reed Solomon coding. Data integrity provided by both fast ECDSA and RS is achieved at the second and third level of the Harricent_RSECC model.

VII. CONCLUSION

In this paper, we appraise that data is vulnerable to attacks and security threats while in transit and/or storage. The study focused on integrating encoding and encryption schemes to offer optimal protection due to unintended data loss as a result of adversary actions or hardware failures. The study implemented a data security model, dubbed Harricent_RSECC data protection model, to improve CAIN of data by dwelling on five key metrics. Through data identification and classification process, the study was scoped with texts and images. The use of RS codes enabled the detection and correction of compromised messages so as to maintain the integrity of the message. To prevent man in the middle attack and message eavesdropping, the model obfuscated the message prior to transmission and storage. Consequently, the confidentiality and authenticity of the message guaranteed. Moreover, through message signing, the model achieved the security levels of integrity and non-repudiation.

VIII. FUTURE WORK

The future study will focus on improving the computational processes so that instead of three deliverables, authors may compress all output into one file. More so, study will expand the scope to cover other scope such audios and videos files.

REFERENCE

- [1] M. M. Kirman, S. M. Saif and A. T. Siddiqui, "Big data : a study of Its issues and challenges," International Journal of Modern Computer Science & Engineering (IJMCSE), vol. 5, no. 1, pp. 29-35, 2016.
- [2] V. Amankona, F. Twum and J. B. Hayfron-Acquah, "A framework for securing data by using elliptic curve cryptography and Reed Solomon coding schemes," in 2021 International Conference on Electrical, Computer and Energy Technologies (ICECET), Cape Town, 2021.
- [3] R. Denis and P. Madhubala, "Hybrid data encryption model integrating multi-objective adaptive genetic algorithm for secure medical data communication over cloud-based healthcare systems," Multimed Tools Applications, p. 21165–21202, 2021.
- [4] D. Sadhukhan, S. Ray, G. Biswas, M. Khan and M. Dasgupta, "A lightweight remote user authentication scheme for IoT communication using elliptic curve cryptography," Journal of Supercomputing, 2020.
- [5] C. A. Lara-Nino, A. Diaz-Perez and M. Morales-Sandoval, "Lightweight elliptic curve cryptography accelerator for internet of things applications," Elsevier: Ad Hoc Networks 103, 2020.
- [6] C. Nakov, "Elliptic curve cryptography (ECC) - practical cryptography for developers," 2019. [Online]. Available: <https://cryptobook.nakov.com/asymmetric-key-ciphers/elliptic-curve-cryptography-ecc>. [Accessed 11 12 2021].
- [7] J. Fadyn, "Basic elliptic curve cryptography using the TI-89 and maple," 26th International Conference on Technology in Collegiate Mathematics., 2015.
- [8] J. Gruska, "Elliptic curves cryptography and factorization," IV054, 2020.
- [9] D. Sadhukhan and S. Ray, "Cryptanalysis of an elliptic curve cryptography based lightweight authentication scheme for smart grid communication," in 2018 4th International Conference on Recent Advances in Information Technology (RAIT), 2018.
- [10] J. Mounika, "Analysis of modified Reed Solomon error correcting codes," International Journal of Recent Scientific Research, vol. 9, no. 6(A), pp. 27225-27228, 2018.
- [11] I. S. Reed and G. Solomon, "Polynomial codes over certain finite fields.," Journal of the Society for Industrial and Applied Mathematics, vol. 8, no. 2, p. 300–304, 1960.
- [12] W. N and J.-Y. C., "Performance Analysis of (255, 239). Reed Solomon Code for Efficient Knowledge-based Systems in Ubiquitous Environment," International Journal of Innovative Technology and Exploring Engineering (IJITEE), vol. 8, no. 852, 2019.
- [13] S. Liu and I. Tjuawinata, "On 2-dimensional insertion-deletion Reed-Solomon codes with optimal asymptotic error-correcting capability," Elsevier: Finite Fields and Their Applications, vol. 73, 2021.
- [14] F. Twum, J. Hayfron-Acquah, W. Oblitey and M.-D. William, "Reed Solomon encoding: simplified explanation for programmers.," International Journal of Computer Science and Information Security (IJCSIS), vol. 14, no. 12, 2016.
- [15] F. Twum, B. J. Hayfron-Acquah, W. W. Oblitey and R. K. Boadi, "A Proposed Algorithm for Generating the Reed-Solomon Encoding Polynomial Coefficients over GF(256) for RS[255,223]8,32.," International Journal of Computer Applications (0975 – 8887), vol. 156, no. 1, 2016.
- [16] W. J. Leis, "Data Transmission and Integrity.," 2018.
- [17] D. He, H. Wang, M. Khan and L. Wang, "Lightweight anonymous key distribution scheme for smart grid using elliptic curve cryptography," IET Communications, vol. 10, no. 14, pp. 1795-1802, 2016.
- [18] E. S. B. Hureib and A. A. Gutub, "Enhancing Medical Data Security via Combining Elliptic Curve Cryptography and Image Steganography," International Journal of Computer Science and Network Security, vol. 20, no. 8, 2020.
- [19] D. S. Laiphrakpam and M. S. Khumanthem, "Image Encryption using Elliptic Curve Cryptography.," Procedia Computer Science, Elsevier, no. 54, p. 472 – 481, 2015b.
- [20] D. S. Laiphrakpam and M. S. Khumanthem, "Implementation of Text Encryption using Elliptic Curve Cryptography.," in Eleventh International Multi-Conference on Information Processing-2015 (IMCIP-2015)., 2015a.
- [21] R. Obaidur, "Data and Information Security in the Modern World by using Elliptic Curve Cryptography.," Computer Science and Engineering, vol. 7, no. 2, 2017.
- [22] A. M. Sauber, P. M. El-Kafrawy, A. F. Shawish, M. A. Amin and I. M. Hagag, "A New Secure Model for Data Protection over Cloud

- Computing," Hindawi. Computational Intelligence and Neuroscience., vol. 2021, p. 11, 2021.
- [23] O. Riznyk, Y. Kynash, O. Povshuk and Y. Noga, "The Method of Encoding Information in the Images Using Numerical Line Bundles," in IEEE CSIT , Lviv, 2018.
 - [24] W. Zhang, S. Zou and Y. Liu, "Iterative Soft Decoding of Reed-Solomon Codes Based on Deep Learning," IEEE Communications Letters, 2020.
 - [25] A. A. Salim and A. B. M. Amal, "Data security for cloud computing based on elliptic curve Integrated encryption scheme (ECIES) and modified identity based cryptography (MIBC)," International Journal of Applied Information Systems (IJ AIS), 2016.
 - [26] A. G. Amar, C. Nouredine and F. Mezrag, "Performance Evaluation and Analysis of Encryption Schemes for Wireless Sensor Networks," IEEE, 2019.
 - [27] M. Dindayal, A. K. Danish and K. Y. Dilip, "Security Analysis of Elliptic Curve Cryptography and RSA," Proceedings of the World Congress on Engineering, vol. 1, 2016.
 - [28] R. S. Soram, K. K. Ajoy and R. S. Soram, "Performance Evaluation of RSA and Elliptic Curve Cryptography.," in 2nd International Conference on Contemporary Computing and Informatics , 2016.

APPENDIX I

A python simulation of the implementation of Harricent_RSECC data protection model:

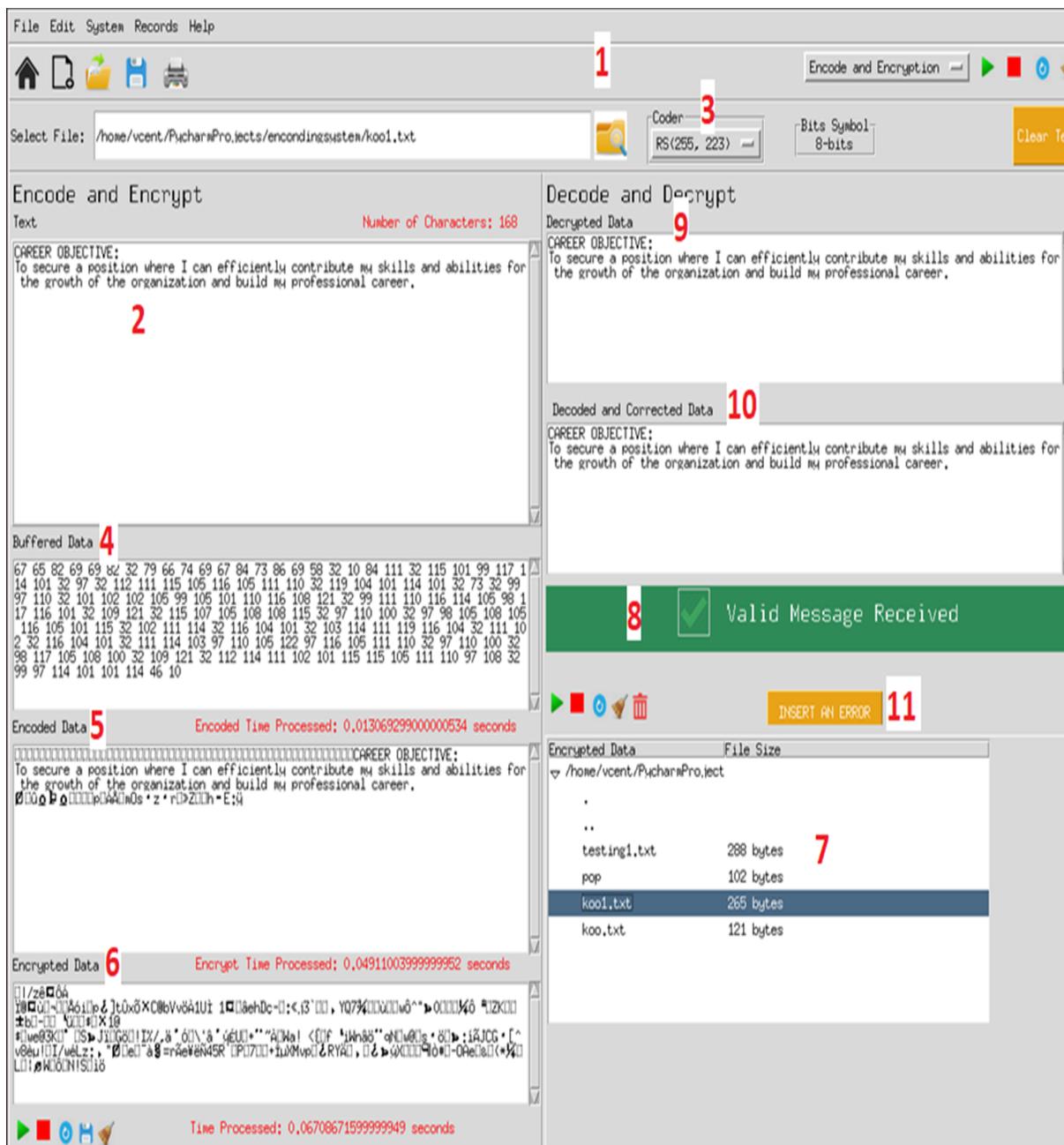


Fig. 9. A Snapshot of the Simulation of the Harricent_RSECC Data Protection Model.

APPENDIX II

Following are some python code snippets for the implementation of Harricent_RSECC Data Protection Model:

```
import unireedsolomon as r
# encoders
rs4a = rs.RSCoder(15, 9, generator=2, prim=0x13, fcr=0, c_exp=4)
rs8d = rs.RSCoder(255, 251, generator=2, prim=0x11d, fcr=0, c_exp=8)
remainderpass = []
codewordpass = []
#the Encoding
if self.codevariable.get() == "RS(53, 37)":
    if thewordlength > 9:
        #messagebox.showinfo(title="Encoding", message="Message length cannot be greater than 37")
        newfxd = self.chunkmydata(fxd, 37)
        fxdlength = len(self.chunkmydata(fxd, 37))
        if fxdlength > 10:
            messagebox.showerror(message="File Size is too big. Please Choose Another Encoder.")
        else:
            for elem, data in enumerate(newfxd):
                codeword = rs8a.encode(data)
                remainder = codeword[-16:]
                print('remainder(hex)= %s' % remainder.encode().hex())
                self.codeworddisplay(codeword)
                remainderpass.append(remainder)
                codewordpass.append(codeword)
            remainderpass.append(remainder)
            codewordpass.append(codeword)
    encoder = "RS(15, 9)"
```

Listing. 1. The Generation of RS Codeword and Remainder.

```
from ecies.utils import generate_eth_key, generate_key
from ecies import encrypt, decrypt
#encryption-key generation
secp_k = generate_key() #random number
sk_bytes = secp_k.secret # generate priv-k using the random key
pk_bytes = secp_k.public_key.format(True) # generate pub-k using the random key
#encrypting data using ECIES
data = bytearray(fxd, "utf8")
encrypted_data = encrypt(pk_bytes, data) #encrypt data using pub-k
self.functiondisplay(newdata, encrypted_data)
```

Listing. 2. The Implementation of the Data Encryption Module using ECIES.

```
from fastecdsa import curve, ecdsa, keys
from fastecdsa.curve import P256
from fastecdsa.keys import export_key, gen_keypair, import_key
from fastecdsa.encoding.der import DEREncoder
from hashlib import sha3_256
#encryption-key generation
# integrity
private_key, public_key = gen_keypair(P256)
# sign
r, s = ecdsa.sign(fxd, private_key, hashfunc=sha3_256)
```

Listing. 3. The Implementation of the Message Signature using Fast ECDSA Algorithm.

Goal Question Metric as an Interdisciplinary Tool for Assessing Mobile Learning Application

Sim Yee Wai¹, Cheah WaiShiang², Piau Phang³, Kai-Chee Lam⁴, Eaqerzilla Phang⁵, Nurfaeza binti Jali⁶

Faculty of Computing and Engineering, Quest International University (QIU), Perak, Malaysia¹

Faculty of Language and Communication, Universiti Malaysia Sarawak, Kota Samarahan, Malaysia⁴

Faculty of Computer Science & IT, Universiti Malaysia Sarawak, Kota Samarahan, Malaysia^{2,3,5,6}

Abstract—Assessing the mobile learning application among interdisciplinary researchers is a non-trivial task. Mandarin Learning App is a Mandarin 3D game tailor-made for students who choose PBC1033 Mandarin Language Level 1 as an elective course. It is an interdisciplinary project which it involves researchers from software engineering, computational science/mathematics and the Faculty of Language studies. In the project, the software engineer focuses on producing a quality application mostly through usability studies; the language teacher focuses on students' study performance upon using the Mandarin Learning App and the mathematician focuses on finding the statistical data dependency of collected data through the various statistical packages. Hence, we are facing issues like how to reach a consensus in working on assessing the Mandarin 3D games? How to enable the discussion among the researchers; how to consolidate the results so that we can understand? We introduce Goal Question Metric to tackle these issues. In this paper, we demonstrate how Goal Question Metric is used to form a holistic view of assessing requirements on mobile applications and guide the discussion and reach consensus in analyzing the results of the evaluation. The contribution of this paper is to introduce Goal Question Metric as an interdisciplinary tool while assessing the mobile learning application. With Goal Question Metric, we demonstrate how it can structure the assessment from a different viewpoint in a comprehensive and systematic manner; 1) better structure of the experiments, 2) able to reach consensus among researchers from different disciplines, 3) able to analyze the dependencies among various experiments and 4) able to find hidden results.

Keywords—Goal question metric; mobile application; evaluation; communication; interdisciplinary

I. INTRODUCTION

Language is very important for everyone to have connection and interaction with others either in verbal or nonverbal communication. UNIMAS has offered introductory Mandarin Language courses for students who wish to explore and learn Mandarin. Among them, the PBC0033 Mandarin Language Level 1 is offered as an elective course to students who do not have any basic knowledge of Mandarin. In this course, students start to learn and recognize the Chinese character; to pronounce the character together with pinyin and to write the character with the correct stroke. Based on our informal observation, students are interested to adopt mobile applications to learn Mandarin. Hence, we developed an in-house mobile application known as "Mandarin App", under an interdisciplinary project between researchers from software engineering, language teacher and mathematician.

In the project, the software engineer focuses on producing a quality application by evaluating the system mostly through usability studies. On the other hand, the language teacher focuses on students' study performance upon using the Mandarin Learning App and the mathematician focuses on finding the statistical dependency of collected data through the various statistical packages. Hence, we are facing issues like how to reach a consensus in working on assessing the Mandarin 3D games? How to enable the discussion among the researchers? How to understand the assessing mechanism in which some are too technical (e.g., statistical analysis) and how to consolidate the results so that we can understand? Hence, we are in the dilemma to reach a consensus when conducting the evaluation of the project.

We introduce Goal Question Metric (GQM) to tackle these issues. In this paper, we demonstrate how Goal Question Metric is used to form a holistic view of assessing requirements on mobile applications and guide the discussion and reach consensus in analyzing the results of the evaluation. The contribution of this paper is to introduce GQM as an interdisciplinary tool while assessing the mobile learning application. With GQM, we demonstrate how it can structure the assessment from a different viewpoint in a comprehensive and systematic manner. Section II presents the related works in using GQM to assess a system. Section III presents how GQM is used to assess a mobile application. Section IV presents the results derived from GQM measurement guidelines. The paper is concluded in Section V.

II. RELATED WORK

GQM is used to perform usability evaluation of real-time water quality monitoring mobile applications [1] and general mobile applications [2], [3], [4], [5], [6], [7], [8], [9]. It describes the data that we need to collect and how to interpret the data [10]. The GQM consists of goal, questions, and metrics. Several metrics are identified in user experiment measurements on mobile applications [11]. They are understandability, learnability, efficiency, effectiveness, operability, attractiveness, usability compliance, happiness, engagement, adoption, retention, task success, usability, satisfaction in use, safety, social presence, cross-platform interaction, algorithm diversity, user control, usage effort, outcome-related experience, review, connection, content, internet service, service availability, ease to use, utility, long term use, productivity, generalizability, pragmatic quality, hedonic quality, stimulation, emotion, psychology metric.

Other metrics are task completion, error rates, time to take the usage and subjective satisfaction like enjoyment, ease of use and safety [1]. Saleh et al. [12] classified the usability metrics into 7 criteria. They are satisfaction, efficiency, effectiveness, learnability, operability, universality, and user interface aesthetics. The satisfaction covers metrics of comfort, trust, pleasure, usefulness; the efficiency covers metrics like task efficiency, time efficiency, relative task time; effectiveness covers metrics of task completion, task effectiveness and effort frequency; the learnability covers metrics of time to learn, memorability, easy to understand error messages, completeness of user documentation, cognitive load; operability covers metrics of understandable, message clarify, operational consistency; universality covers metrics of cultural universality, standard compliance, accessibility; user interface aesthetics covers metric of customizability and attractive of the user interface.

The GQM is extensible to serve as a measurement guideline to measure mobile applications [1]. For example, [1] extend the GQM with guidelines to determine how a specific goal is reached. The guideline can drive the formation of questions for GQM. Meanwhile, the guideline will serve as expected answers for the metrics. For example, to achieve goal simplicity, a system should be easy to input the data, easy to install, easy to learn. Hence, the evaluator can form the questions like is it simple to key in data? How easy it is to install the application? etc. In addition, GQM [10] has been used in running a course activity to train the student's ability to understand its goal, refine its goal and design appropriate achievement metrics.

We extend the usage of GQM from usability evaluation into a wider assessment mechanism. Hence, GQM for mobile application evaluation should not only focus on usability study of mobile applications but it can also be used and support various experiments analysis. This is in line with the work [1] to extend the usage of GQM with guidelines. The guideline presents the expected answers for the metrics. On the other hand, [13] extended the GQM with instrument types and role mapping to relate the evaluation and analysis methods into metrics. Meanwhile, our work is related to [10] by treating GQM as a tool for stakeholders during the assessment of mobile learning application.

III. THE METHOD

This section presents the adoption of GQM as an interdisciplinary tool to assess mandarin learning app among different researchers.

GQM is known as Goal, Question, Metric [14]. The GQM model is represented as a hierarchical structure and covers three levels. They are conceptual level, operation level and quantitative level. The conceptual level defines the goal of the measurement. It represents the measurement from a high level of abstraction. The goal at this level covers the quality level of the system. Hereby, we treat is quality goal. The quality goal is to act as a non-functional goal and represents how a functional goal should have been achieved. The quality goal is the property of the system. Quality goals are non-functional goals (sometimes referred to as soft goals). The operation level defines a set of questions that will work towards the

assessment or achievement of a specific goal. The quantitative level defines a set of data that serves as the answers to the questions. The data can be objective or subjective depending on the measurement mechanisms (e.g. quantitative or qualitative assessment method).

The GQM processes cover the process to determine what to measure and determine how to measure. To determine what to measure, it involves processes to identifying entities, classifying entities to be examined and determine the relevant goals. Meanwhile, it is important to inquire about metrics and assign the metrics to determine how to measure.

In this section, we introduce how to adopt GQM as an interdisciplinary tool to assess mandarin learning apps among researchers from different disciplines.

The goal is the first step when working on the GQM model. The goal represents the objective or purpose of a system, and it is achievable [15]. Since GQM is used to measure the system, the goal must relate to what is the purposes to measure the entities regarding the issues been solved by whom. The goal is related to software quality characteristics as described in [1]. The goal here also refers to what do you want to know or learn from the system. The formation of the goal is based on the template given in [16]. According to [16], a goal is derived based on the elements like purpose, issues, object and viewpoint [16]. Goals need to be high level. We can derive the goals by referring to the documents' analysis, interviews [14] or existing usability models as listed in [1], [13], [12], [2], [17].

Questions are elicitation questions in achieving the goal as stated previously. The generation of the questions is based on the instrument decided and targets users.

Metrics are the measurement parameters in answering the questions towards achieving the goal. Metrics produces data. The data can be objective or subjective depending on the types of measurement mechanisms. For example, a quantitative analysis leads to objective data (e.g. number of satisfaction; the number of scores). Meanwhile, the qualitative analysis led to subjective data. Once, the GQM is formed, a goal-question-metrics refinement process takes place to enable the researchers to reformulate the goal, questions, and metrics on evaluating the mandarin learning app.

As mentioned before, this is an interdisciplinary project. The software engineer has the intention to develop a quality product and usability studies are among the famous technique for assessing it. On the other hand, language teacher is always focuses on study performance. This is done through pre-test and post-test. In addition, an empirical study is used to study the correlation of results regarding demographic and various post-survey as proposed by the mathematician. Hence, how to consolidate the different types of analysis and studies?

To evaluate the effectiveness of the games, several experiments are conducted. Data are collected by using both the written questionnaire surveys and interviews, a vocabulary pre-test and post-test (before-after), as well as game diaries. Collected data were analyzed with quantitative and qualitative methods. Specifically, exploratory factor analysis was performed to reexamine the grouping of variables in

questionnaire items to establish underlying dimensions that could explain its correlations.

All the experiments are conducted among 33 students from PBC0033, the batch year 2021 who answered the questionnaire and took the pre-test and post-test. As our sample size is very small, we only remove 3 respondents who rate their responses on Likert-type questionnaire items with the same answer always, resulting in a standard deviation lower than or equal to 0.385. Hence, the characteristics and demographics of our remaining 30 respondents are summarized in Table I.

TABLE I. RESPONDENT DEMOGRAPHICS AND CHARACTERISTICS

Demographic or characteristic	Frequency	Per cent
Gender:		
Male	5	16.5
Female	25	83.3
Ethnic:		
Bumiputera	28	93.3
Dusun	2	6.7
Age:		
21 - 22	20	66.7
23 - 24	8	26.7
25 - 26	2	0.6

Prior to the evaluation, the GQM is adopted and serve as a tool to drive the discussion among the researchers. Based on GQM, Table II shows the list of goals, questions, and metrics in related to our study. Meanwhile, Fig. 1 shows the GQM model in assessing the mobile learning App. It starts with a higher-level goal to ensure the effectiveness of learning through the mobile Mandarin App.

To achieve the main goal, there exists sub-goals of 'ensure engaging', 'ensure high performance', 'ensure likelihood', 'ensure secure learning', 'ensure privacy protection' and 'ensure cross culture learning'. The goals are derived from the Table I. To achieve the subgoals, several questions are generated and the answers from the questions will lead to the achievement of the sub-goals. In this case, the answer from "Do the students like to use the app to learn" will achieve the goal to ensure engaging; the answers from "are the games able to improve students' results" will achieve the goal of ensuring high performance; the answer from 'what are the factors that influence the adoption of games' will achieve three subgoals to ensure likelihood of the learning. In addition, we believe that the answers also lead to achievement of subgoals of namely ensure secure learning, ensure privacy protection, ensure cross culture learning after postmortem. The metrics that are corresponded to the questions are level of preference, level of self-learning, level of games challenges, level of critical thinking; number of pre-test level, number of post-test level, comparison of the results; number of dimensions, list of constructs; number of incidents and amount of lost. Finally, we have mapped the metrics to the testing instruments and the researchers that are handled or initiated by the researchers. The mapping of the metrics into instruments is presented in [13].

In sum, the GQM model can serve as a communication media to discuss among the researchers. Before this, we only identified three goals for the assessment. They are ensuring engaging, ensure high performance and ensure likelihood.

After the postmortem, three subgoals have been identified and three questions are derived. They are ensuring secure learning, ensure privacy protection and ensure cross culture learning although it does not cover in this evaluation. It shows how GQM model can be used to find the hidden evaluation elements during the discussion. In addition, the corresponding questions are 'Does the games secure?' 'Do the games able to protect privacy?' and 'Does the games able to promote cross culture learning?'

TABLE II. THE GQM LIST FOR THE STUDY

GQM		Item
Main Goal	Purpose	Improve
	Issue	Learning of mandarin through
	Object (process) Viewpoint	Mobile learning app Among students
Question		Does it an effective way to use mobile learning app?
Metric		Refer to the rest of the description.
Goal 1	Purpose	Ensure
	Issue	engaging in
	Object (process) Viewpoint	learning through mobile app among students
Question		Do the students like to use the app to learn Mandarin?
Metric		Level of preferences to use mobile games Level of self-learning Level of games challenges Level of critical thinking
Goal2	Purpose	Ensure
	Issue	high result performance
	Object (process) Viewpoint	by learning through mobile app among students
Question		Are the games able to improve student results?
Metric		Number of pre-test level Number of post-test level Comparison of the results
Goal3	Purpose	Ensure
	Issue	likelihood of student behaviour
	Object (process) Viewpoint	by learning through mobile app among students
Question		What are the factors that influences the adoption of games among students?
Metric		Number of grouping/ numbers of dimension List of constructs

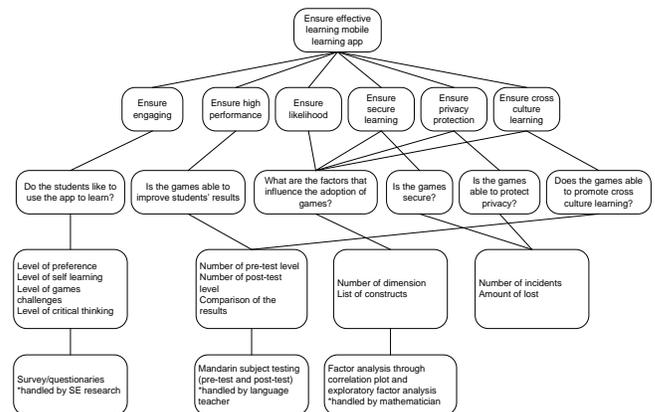


Fig. 1. GQM Model for the Project's Evaluation.

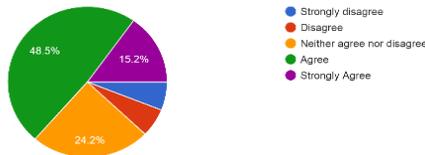
IV. THE RESULTS

A. Results for Goal 1

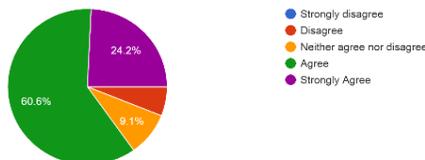
- Question: Is the games able to become a complemented tool on conventional Mandarin lesson?
- Answer: From the results, the mandarin learning app can serve as a complemented tool on conventional mandarin lesson.

Questionnaires are used as the method to achieve the goal. We have formulated the questionnaire based on the metrics given. The results from the usability study are shown in Fig. 2. The survey is designed based on five elements. They are students' motivation; students' attitudes with the games; students' cognitive development; games interface; students' expectation.

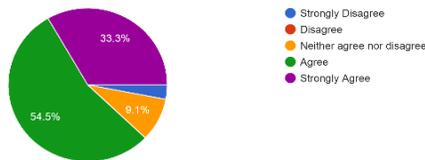
2. I prefer to answer questions using mobile games as compared to using books or paper. 33 responses



3. It is more flexible for me to determine my own learning time. 33 responses



5. These games challenge my understanding of the subject. 33 responses



1. These computer games help me to think critically. 33 responses

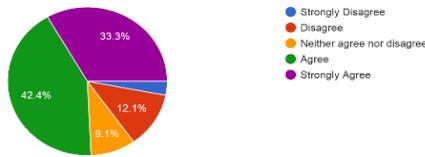


Fig. 2. Usability Study Results.

In sum, students are motivated to use the games in learning Mandarin. 65% of the students are interested to use the games in learning Mandarin instead of using books and paper. The games increase the student learning in which students can learn anytime and anyway (85%). The games can challenge the students to think critically and enhance problem solving in which 76% of respondents stated. Meanwhile, 87% of the

respondents agreed that the games can challenge the students' understanding of Mandarin. It has been reported that the games are interactive and interesting. 79% of the respondents agreed that the menus in the games are easy to understand. 64% of the respondents stated that the navigation and interaction are easy to use. The games are easy to function with a short time learning curve. 91% of the respondents stated that the multimedia elements in the games are interesting. 82% of the respondents show the interest to replicate the games mode in other level of Mandarin course. Although games are interesting, 50% of the respondents still prefer to have the traditional face to face class. This serves the objective on having the games as a complemented tool on conventional Mandarin lesson.

B. Results for Goal 2

- Question: Is the games able to improve student results?
- Answer: The post test results are more than pre- test results. Hence, the games can improve student results.

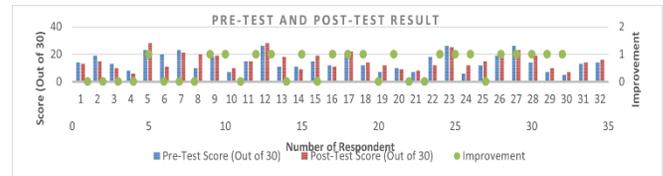


Fig. 3. Individual Pre-test and Post-test Score.

Pre-test and post-test are adopted in this study. We first give the individual pre-test and post-test scores in Fig. 3. Only slightly more than half of the students showed improvement in their test scores after going through mobile games enhanced learning. We then check the normality of the pre- and post-test scores by using Shapiro-Wilk test. Both the test scores are normally distributed as their p-values > 0.05. From their boxplots (see Fig. 4), we find that there is only a marginal improvement whereby the means of pre-test and post-test score are 14.4 and 15.4, respectively. Although the paired sample t-test supports that the true difference in means between pre- and post-test score is not equal to 0, if we plot the change scores (i.e. post-test minus pre-test scores) against pre-test scores, as shown in Fig. 5, a negative correlation can be observed. This implies that a common statistical phenomenon in repeated measurements, known as regression towards the mean (RTM) [20], can influence the findings from our pre- and post-test instruments. This RTM suggests that students with higher pre-test scores consistently make smaller improvement in post-test than students with lower pre-test scores.

As we are not able to use a randomized control group to deal with this RTM, we opt to assess the agreement between these two pre- and post-test measurements through Bland-Altman plot (see Fig. 6). The black line gives the mean of difference in score between pre- and post-test while the two blue dashed lines represents the 95% confidence interval limits of agreement for the mean of difference. In our example, the mean of difference is 0.96. This suggests that on average the post-test measures 0.96 score more than the pre-test as mean of difference (called bias) is non-zero. Besides, only two points appear to lie outside the limits of agreement indicating there is certain degree of agreement between the two tests.

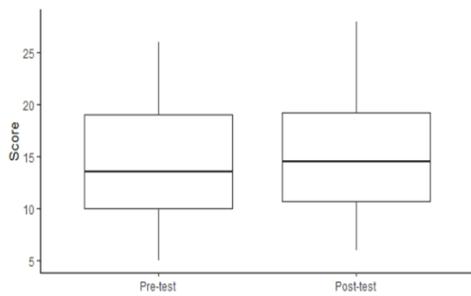


Fig. 4. The Boxplots for Pre- and Post-test Score.

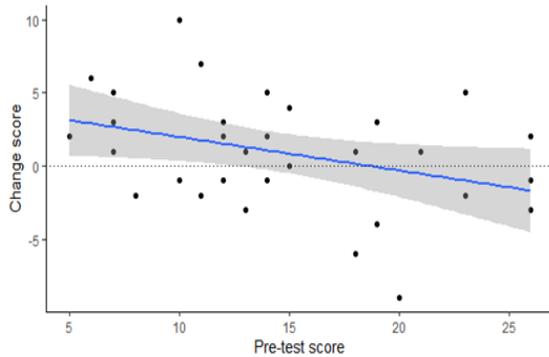


Fig. 5. Scatter Plot showing Change Score against Pre-test Score.

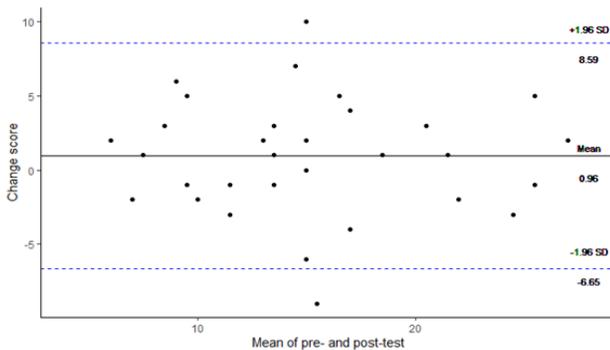


Fig. 6. Bland-Altman plot with the Representation of the Limits of Agreement (Dotted Line).

C. Results for Goal 3

- Question: What are the factors that influence the adoption of games among students?
- Answer: They are two major dimensions in making the mandarin learning app success among our students. The success of the mandarin learning app is due to high student motivation, high cognitive development from the games plays and the games meet the student expectation. Meanwhile, the games are user friendly.

A set of questionnaires with 5 constructs (motivation, attitudes, cognitive development, interface design and expectation) with 24 items and an open-ended question on student’s open comments about learning using the games is designed as shown in Table III. We present the correlation plot and exploratory factor analysis for questionnaire items in all 5 constructs to answer derived on this goal.

TABLE III. THE DETAILED OF QUESTIONNAIRE WITH 5 CONSTRUCTS

Section	Item	Question
Student’s motivation	A1	I think this gaming activity gives me lots of benefits.
	A2	I prefer to answer questions using mobile games as compared to using books or paper.
	A3	I am very interested in using games for learning new Mandarin words in the future.
	A4	Digital learning games do not bring additional value to Mandarin language learning
	A5	I prefer to do exercises in games rather than quizzes during class.
	A6	The usage of computer games makes this subject more interesting.
Student’s attitudes with the games	B1	With the games, I can learn better by myself.
	B2	I can learn better according to my own pace and sequence.
	B3	It is more flexible for me to determine my own learning time.
	B4	It is more flexible for me to choose my learning pace.
	B5	The content of the games matches my subject syllabus.
	B6	The element of multiplayer in games motivates me to study Mandarin characters.
	B7	The games do not affect my motivation to learn new words.
Students’ cognitive development	C1	These computer games help me to think critically.
	C2	Solving the given problems is very interesting.
	C3	It is worth using games for learning in future.
	C4	Looking for the answer to questions given is an encouraging activity.
	C5	These games challenge my understanding of the subject.
Game interface	D1	Menus available in the games are easy to understand.
	D2	Navigations and interactions are easy to use.
	D3	Multimedia elements in the games are interesting.
	D4	I just need a very short time to know how the game is functioning.
	D5	The use of colour and design layout in the games are interesting.
Students expectation	E1	I wish I have more opportunities to learn other Mandarin course using this game approach.
	E2	I prefer using games to learn Mandarin as compared to traditional methods in the class.
	E3	I would like to learn all computer subjects using educational games.
	E4	I wish this game will be available online for easy access.

We carried out the analysis using RStudio, specifically by using package ‘psych’ [18]. We first reverse the Likert score for variables A4 and B7 as they are negatively phrased items (hereinafter, the terms “variable” and “item” are used interchangeably). As Likert-scale questionnaires yield ordinal data, we measure the association of the ordinal variables in terms of polychoric correlation and estimate the internal

consistency for scales using ordinal alpha. The coefficient ordinal alpha gives 0.94 indicating a very good level of reliability. Furthermore, this alpha value does not increase a lot when any of the items is deleted. Therefore, all the items as given in Table III are considered for further analysis.

We explore the strength of the relationship between all variables by constructing correlation plot with package ‘corrplot’ [19], as shown in Fig. 7. The correlation is positive (resp. negative) when one variable increases as the other increases (resp. decreases). We find that the two negatively phrased items (i.e. A4 and B7) as well as item D4 have negative correlation with all other variables, suggesting that item D4 probably should be reversed. However, we chose not to reverse item D4 scoring as it showed that our respondents understood the questions correctly.

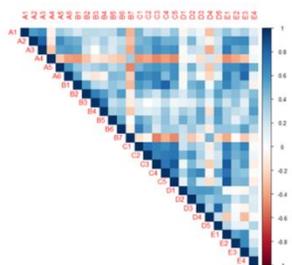


Fig. 7. Correlation Plot for all Likert-type Questionnaire Items.

We then employed multivariate statistical technique to further identify the possible latent relational structure among all the variables. Particularly, we performed exploratory factor analysis (EFA) to reexamine the grouping of variables and establish underlying dimensions that could explain the correlations, thereby allowing the formation and refinement of theory in the context of using mobile games to support Mandarin learning to a group of non-native learners. In other words, we tested the construct (i.e. factor) validity of the questionnaire with the factor analysis. The factors that explain the highest proportion of variance the variables share is expected to represent the underlying constructs [21].

To determine the suitability of our Likert-type questionnaire data for EFA, we perform Barlett’s test, and it returns p-value < 0.05 showing that the items correlate anyhow and EFA may be useful. Besides, the positive determinant of the correlation matrix suggests that EFA will probably run. We then determine the number of factors to retain by constructing the screen plot, as illustrated in Fig. 8. Together with parallel analysis, we take the number of factors as two for further analysis.

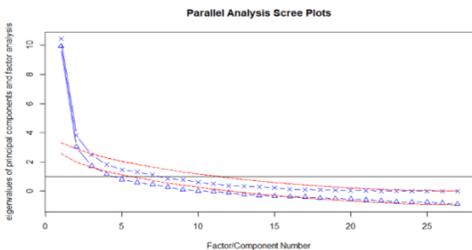


Fig. 8. The Screen Plot based on the Unreduced Correlation Matrix and Parallel Analysis.

We then carried out the exploratory factor analysis using maximum likelihood estimation procedure and its results are depicted in Fig. 9. Seventeen Likert-type items in the questionnaire make up the first factor and 9 items the second factor. Item E4 may be ignored as its load is lower than 0.3 for both factors. This implies that the measurement on students’ expectation is irrespective of whether the mobile game is available online or not. Also, from such a regrouping of variables, we could say that the correlations among all the variables might be categorized into two broad dimensions, in which the first dimension includes the students’ motivation, cognitive development and expectation whereas the second dimension mainly covers game interface. It is interesting to point out that variables related to students’ attitude with the mobile game are split into two dimensions, whereby items B1, B6 and B7 fall under first dimension while the remaining items B2 to B5 are in second dimension.

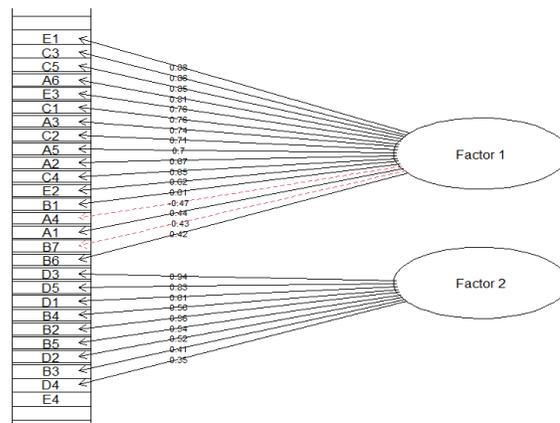


Fig. 9. Regrouping of Variables using Exploratory Factor Analysis.

V. CONCLUSION

This paper presents the adoption of GQM as an interdisciplinary tool to access the mobile mandarin app. With the GQM, we can summary the benefits of GQM as following. 1) Better structure of the experiments; 2) able to reach consensus among researchers from different disciplines; 3) able to analysis the dependencies among various experiments; 4) able to find hidden results. The GQM can structure the experiments and drive the communication among the researchers. It is also able to consolidate various experiments and preset the results in a more systematic manner as described in the previous section. In addition, the GQM can reduce the overlapping of method used, parameters and questions among different evaluation techniques that are used by the researchers. With GQM, we can discover the relationship across the experiments that are conducted by researchers. For example, the pre-test and post test results are able to continue study the cross culture learning among the students. The results can provide the answer to the question in achieving the goal of 'ensure cross culture learning'. On the other hand, the number of dimension and list of constructs can further be extended with factors like security and privacy to ensure the achievement of subgoals, 'ensure secure learning' and 'ensure privacy protection'. The contribution of this research is to demonstrate how GQM can be used as an interdisciplinary tool when assessing mobile learning application among researchers

from software engineering SE, language study, and mathematics. From the results, the mandarin learning app can serve as a complemented tool on conventional mandarin lesson. The post test results are more than pre- test results. Hence, the games can improve student results. They are two major dimensions in making the mandarin learning app success among our students. The success of the mandarin learning app is due to high student motivation, high cognitive development from the games plays and the games meet the student expectation. Meanwhile, the games are user friendly. In future, more works are required to investigate the adoption of others concept from agent models [22][23][24] into GQM. For example, to use HOMER questions in forming the questions is yet to explore.

ACKNOWLEDGMENT

The authors would like to thank UNIMAS to support on this research. This research is under UNIMAS Teaching and Learning grant, SoTL SoTL/FSKTM/2020(1)/002.

REFERENCES

- [1] Bokingito Jr, P. B., and Caparida, L. T., "Usability evaluation of a real-time water quality monitoring mobile application," *Procedia Computer Science*, 197, pp. 642-649, 2022.
- [2] Saleh, A., Ismail, R., and Fabil, N., "Evaluating usability for mobile application: a MAUEM approach," in *Proc. of the 2017 Int. Conf. on Software and e-Business*, pp. 71-77, Dec. 2017.
- [3] Kureerung, P., and Ramingwong, L., "Factors supporting user interface design of mobile government application," in *Proc. of the 2019 2nd Int. Conf. on Information Science and Systems*, pp. 115-119, Mar. 2019.
- [4] Gharaat, M., Sharbaf, M., Zamani, B., and Hamou-Lhadj, A., "ALBA: a model-driven framework for the automatic generation of android location-based apps," *Automated Software Engineering*, 28(1), pp. 1-45, 2021.
- [5] Hashim, N. L., and Isse, A. J., "Usability evaluation metrics of tourism mobile applications," *J. of Software Eng. and App.*, 12(7), pp. 267-277, 2019.
- [6] Noei, E., Zhang, F., Wang, S., and Zou, Y., "Towards prioritizing user-related issue reports of mobile applications," *Empirical Software Eng.*, 24(4), pp. 1964-1996, 2019.
- [7] Hussain, A. and Mohamed Omar, A., "Usability Evaluation Model for Mobile Visually Impaired Applications," *Int. Assoc. of Online Eng.*, 2020. [Online]. Available: <https://www.learntechlib.org/p/216427/> . Accessed on: Feb. 14, 2022.
- [8] Schobel, J., Pryss, R., Schlee, W., Probst, T., Gebhardt, D., Schickler, M., and Reichert, M., "Development of mobile data collection applications by domain experts: Experimental results from a usability study," in *Int. Conf. on Adv. Information Systems Eng.*, Springer, Cham, pp. 60-75, June 2019.
- [9] Tabora Mosquera, J. P., Arango-Lopez, J., Gutierrez Vela, F. L., Collazos, C., and Moreira, F., "Analyzing effectiveness and fun through metrics applied to pervasive gaming experiences," *Universal Access in the Information Society*, 20(3), pp. 545-554, 2021.
- [10] Khomyakov, I., Masyagin, S., and Succi, G., "Experience of Mixed Learning Strategies in Teaching Lean Software Development to Third Year Undergraduate Students," in *Int. Workshop on Frontiers in Software Eng. Education*, Springer, Cham, pp. 42-59, Nov. 2019.
- [11] Arifin, Y., Sastria, T. G., and Barlian, E., "User experience metric for augmented reality application: a review," *Procedia Computer Science*, 135, pp. 648-656, 2018.
- [12] Saleh, M. D. R. B., and Majrashi, K., "An enhanced usability model for mobile health application," *Int. Journal of Computer Science and Information Security (IJCSIS)*, 17(2), 2019.
- [13] Tahir, R., and Arif, F., "A measurement model based on usability metrics for mobile learning user interface for children," *The Int. Journal of E-Learning and Educational Technologies in the Digital Media*, 1(1), pp. 16-31, 2015.
- [14] Carvalho, R. M., de Castro Andrade, R. M., and de Oliveira, K. M., "AQUARIUM-A suite of software measures for HCI quality evaluation of ubiquitous mobile applications," *J. of Systems and Software*, 136, pp. 101-136, 2018.
- [15] Shiang, C. W., Halin, A. A., Lu, M., and CheeWhye, G., "Long Lamai Community ICT4D E-Commerce System Modelling: An Agent Oriented Role-Based Approach," *The Electronic J. of Information Systems in Developing Countries*, 75(1), pp. 1-22, 2016.
- [16] Alqahtani, A., Solaiman, E., Patel, P., Dustdar, S., and Ranjan, R., "Service level agreement specification for end-to-end IoT application ecosystems," *Software: Practice and Experience*, 49(12), pp. 1689-1711, 2019.
- [17] Ngadiman, N., Sulaiman, S., Idris, N., Mohamed, H., and Osman, S. M., "A Survey on Quality Factors in Designing Educational Applications for Active Learning," in *2019 IEEE 9th Int. Conf. on System Engineering and Technology (ICSET)*, IEEE, pp. 23-28, Oct. 2019.
- [18] Revelle, W., and Revelle, M. W., "Package 'psych'," *The Comprehensive R Archive Network*, pp. 338-338, 2015.
- [19] Wei, T., Simko, V., Levy, M., Xie, Y., Jin, Y., and Zemla, J., "Package 'corrplot'," *Statistician*, 56(316), 2017.
- [20] Barnett, A. G., Van Der Pols, J. C., and Dobson, A. J., "Regression to the mean: what it is and how to deal with it," *Int. J. of Epidemiology*, 34(1), pp. 215-220, 2015.
- [21] Taherdoost, H., Sahibuddin, S., and Jalaliyoon, N., "Exploratory factor analysis: concepts and theory," *Adv. in Applied and Pure Mathematics*, 27, WSEAS, pp. 375-382, 2014.
- [22] Cheah, W., ChinHong, P., Halim, A. A., "Agent-Oriented Requirement Engineering for Mobile Application Development," *International Journal of Interactive Mobile Technologies*, 11(6), 2017.
- [23] Zulkifli, S. F. B., Waishiang, C., Khairuddin, M. A. B., Jali, N. B., & Bujang, Y. R. B., "How to Model an Engaging Online Quiz? The Emotion Modeling Approach," *Journal of Telecommunications and Information Technology*, (1), pp. 54-63, 2022.
- [24] Ten LiBin, M., WaiShiang, C., Khairuddin, M. A. B., Mit, E., Erianda, A., "Agent-Oriented Modelling for Blockchain Application Development: Feasibility Study," *JOIV: International Journal on Informatics Visualization*, 5(3), pp. 248-255, 2021.

Spatial Feature Fusion for Biomedical Image Classification based on Ensemble Deep CNN and Transfer Learning

Sanskruti Patel, Rachana Patel, Nilay Ganatra, Atul Patel
Smt. Chandaben Mohanbhai Patel Institute of Computer Applications
Charotar University of Science and Technology, Changa, India

Abstract—Biomedical imaging is a rapidly evolving field that covers different types of imaging techniques which are used for diagnostic and therapeutic purposes. It plays a vital role in diagnosis and treating health conditions of human body. Classification of different imaging modalities plays a vital role in terms of providing better care and treatment options to the patients. Advancements in technology open up the new doors for medical professionals and this involves deep learning methods for automatic image classification. Convolutional neural network (CNN) is a special class of deep learning that is applied to visual imagery. In this paper, a novel spatial feature fusion based deep CNN is proposed for classification of microscopic peripheral blood cell images. In this work, multiple transfer learning features are extracted through four pre-trained CNN architectures namely VGG19, ResNet50, MobileNetV2 and DenseNet169. These features are fused into a generalized feature space that increases the classification accuracy. The dataset considered for the experiment contains 17902 microscopic images that are categorized into 8 distinct classes. The result shows that the proposed CNN model with fusion of multiple transfer learning features outperforms the individual pre-trained CNN model. The proposed model achieved 96.10% accuracy, 96.55% F1-score, 96.40% Precision and 96.70% Recall values.

Keywords—Biomedical images; convolutional neural network; ensemble deep learning; feature fusion

I. INTRODUCTION

Biomedical imaging refers to capturing of an organ or tissue for diagnostic purpose. The field is very broad and rapidly evolving that covers different types of imaging modalities like ultrasound, magnetic resonance imaging (MRI), computerized tomography (CT), positron emission tomography (PET), etc. [1]. Biomedical and medical imaging plays a significant role in diagnosis and treating health conditions of human body. It helps to identify problematic health conditions in their early stages that certainly lead for providing better treatment to the patients [2]. The structural and functional changes in biological tissues of the human body normally cause the possible health problems. Biomedical imaging provides a way to view inside the human body that helps to reveal such changes [3].

Classification of different imaging modalities plays a vital role in terms of providing better care and treatment options to the patients. The traditional way to classify these different modalities of images is the naked eye classification that is

performed by medical professional or subject expert. This is the cumbersome and sometimes time-consuming method. Advancements in technology open up the new doors for medical professionals that involve computer-aided diagnosis (CAD) methods for automatic image classification [4]. The increasing advancement in the field of medical imaging technology, medical research and diagnosis become easy. Various types of imaging modalities and procedures are included in medical imaging technology that helps in diagnosis and treatment of the patients. Hence, it plays a dominant role in deciding the actions for the benefit of the patient's health.

In past few years, artificial intelligence (AI) brings a new way for analysing and interpreting data that also called predictive analysis that helps to identify the early signs of any of the health conditions [5]. Deep learning, an approach of AI, emerged as an outperformer in interpreting and analysing image data. Significant advancement has been made in the field of medical image diagnosis that improves disease diagnosis process significantly. Deep learning uses the architecture of artificial neural network that mimics the working of human brain. The complex computer vision tasks are effectively solved by the deep learning algorithms such as image recognition, classification and segmentation. The special class of deep learning algorithms is known as a convolutional neural network (CNN) is widely used to solve image classification problems and achieved a significant performance on benchmark datasets. The reason behind popularity of CNN is the large availabilities of datasets and support of powerful Graphics Processing Units (GPU) that makes the integration of deep learning methods with computer vision popular [6].

There are several distinct imaging modalities in which biomedical images are generated. They are different in shape and types. Due to the diverse data distribution patterns, it may happen that the same CNN model may show different performance on different datasets [7]. CNN models are sensitive to the particulars of the training data. This makes possible that each time they are trained; they may find a different set of weights. These different predictions generate high variance [8]. Moreover, these deep features face the problem of small intra-class variance and large inter-class variance [9].

To address the above issue, a novel spatial feature fusion-based approach for biomedical image classification based on ensemble Deep CNN and transfer learning is proposed. We

have used stacked generalization that is an ensemble learning method. We have used four benchmarked pre-trained CNN models as a deep feature extractor. Extracted features from different networks are merged using spatial feature fusion method. Finally, two FC (fully connected) layers with ReLU are used along with one FC layer with softmax activation function.

The major contributions of the paper include: (a) To propose an ensemble learning framework for creating a deep feature extractor that combines transfer learning features from more than one pre-trained CNN models (b) To apply a spatial feature fusion technique that creates a generalized feature space from extracted features (c) To proposed a deep CNN model that can be used for biomedical image classification with increased prediction accuracy.

II. RELATED WORK

Health care is one of the fastest growing sectors that is delivered by health experts. It is sometimes difficult to recognize disease patterns from huge number of medical images. The special classes of Artificial Intelligence (AI) are machine learning and deep learning algorithms. These algorithms give impressive results for classification of biomedical images. Image classification is considered as one of the computer visions tasks. Many researchers have been worked in biomedical classification that resulting several robust methods that can be categorized into two types; traditional digital image processing techniques and deep learning models. The traditional image processing methods involves manual feature extractions methods whereas deep learning models perform the feature extraction without manual intervention.

N. Sharma et al [10] used texture primitive features to perform segmentation along with classification of medical images. They have used artificial neural network (ANN) for designing the algorithm that performs segmentation and classification. The algorithm has been used for CT and MRI images of distinctive body parts like brain and liver. Ahmed M. Sayed [11] applied machine learning techniques for diagnosis of breast cancer from MRI images. The highest classification accuracy is 94/6% given by KNN. M. I. Daoud et al [12] proposed a fusion approach based on multiple ROI for classification of breast ultrasound image. GLCM texture features are extracted in each ROI that are further classified using binary SVM classifier. The dataset considered for the experiment contains 64 benign and 46 malignant tumor images. The proposed approach provides very promising results with 98.2%, 98.4%, and 97.8% values for accuracy, specificity, and sensitivity respectively. P. Chak et al [13] proposed an Artificial Neural Network (ANN) and SVM based approach to classify the kidney stone images. To extract the features from the CT images, GLCM method was used. The ANN approach gives 95% accuracy, whereas SVM approach gives 99% accuracy. P. Nanglia et al [14] proposed a hybrid algorithm for classification of lung cancer images. The dataset considered for

an experiment contains 500 images and the overall accuracy of the approach is 98.08%.

Deep learning is effectively applied for various domains, including satellite imaging, observation frameworks, mechanical and medical procedures, and precision agriculture. Several researchers have worked upon applying deep convolutional neural network for different medical applications. H. T. Nguyen et al [15] improved the prediction of disease using shallow CNN. They have applied data visualization techniques on Metagenomic data and achieved promising results. S. M. Anwar et al [16] proposed deep transfer learning and LDA (Linear Discriminant Analysis) based approach for classification of medical image modality. They have considered pre-trained ResNet-50 model to implement transfer learning approach along with LDA approach. For experiment, a benchmark ImageCLEF-2012 dataset was considered. The classification accuracy obtained is 87.91% that is significantly better as compared to the state-of-the-art approaches. C.-H. Chiang et al [17] applied CNN for automatic classification of medical image modality. They have considered multiple image modalities that include CT images of abdomen & brain and MRI images of brain and spine. The accuracy achieved for validation and test sets are greater than 99.5%. Moreover, the F1-score for each of the category is greater than 99%.

B. P. Battula and D. Balaganesh [18] propose a hybrid model for medical image classification based on CNN and Encoder. HIS2828 and ISIC2017 are the datasets considered for the experiment. The results show that the accuracy of the proposed model is better than the existing models. A. A. Goma et al [19] discussed about how CNN and GANs are used to improve early prediction of plant disease. They have considered a tomato plant images that are infected with Tomato Mosaic Virus for conducting an experiment. The proposed CNN provides 97% accuracy. S. Patel [20] classified bacterial colony images using Atrous convolution using transfer learning approach. The dataset considered for the experiment contains 660 bacterial colony images classified into 33 distinct classes. The proposed model replaces the standard convolution layer of the VGG-16 pre-trained model with the Atrous convolution. The training and validation accuracy obtained from the experiment are 95.06% and 93.38% respectively.

III. PROPOSED METHOD

In this research, we have proposed a novel spatial feature fusion based Deep CNN model for biomedical image classification. We have developed a feature fusion network using Ensemble deep CNN that leverages the power of state-of-the-art pre-trained CNNs. Ensemble learning is the technique that combines different individual CNN models in order to increase prediction accuracy through generalization [21]. As shown in Fig. 1, the proposed model has three stages, (a) deep feature extraction and feature maps (b) spatial feature fusion of deep transfer learning features and (c) classification.

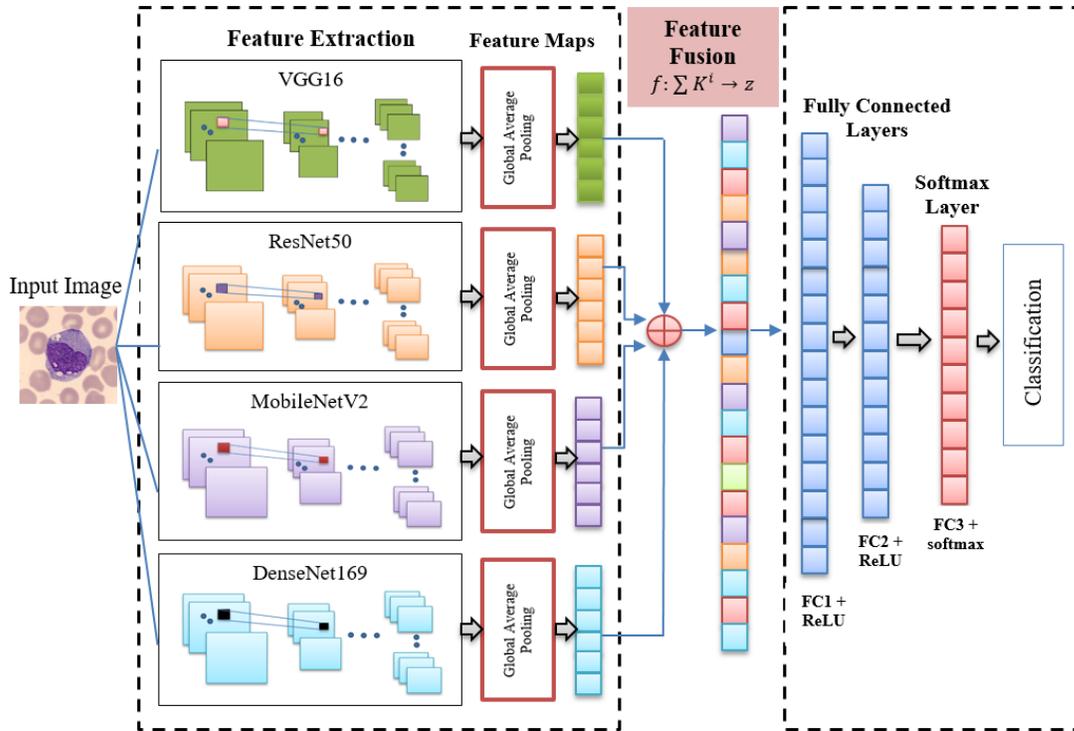


Fig. 1. Spatial Feature Fusion based Deep CNN Model using Ensemble Deep CNN and Transfer Learning.

A. Deep Feature Extraction and Feature Maps

Feature extraction and feature map generation is the crucial step for classification of images using deep learning networks. A deep learning network normally requires a large amount of resources and dataset to be trained for precise feature extraction. The common practice followed is to use the pre-trained model instead of building and training the CNN model scratch. There are several pre-trained CNN models already exist which are developed to solve a similar problem [22]. All modern pre-trained CNN models have emerged from the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) challenge [23]. The challenge is organized yearly from 2010 to 2017 that provides a platform for researchers to present their algorithms related to object localization and image classification tasks. The pre-trained models belong to ILSVRC are trained on ImageNet dataset that consists about 14 million images categorized in several classes. The pre-trained models used the concept of transfer learning, i.e. knowledge gained by solving one problem can be applied to solve similar related problems by using an already trained model to learn a different set of features. It eliminates the need of large amount of data with significant better performance along with reduced convergence time [24]. The commonly used pre-trained models are VGG16, VGG19, MobileNet, InceptionNet, ResNet, DenseNet, etc.

We have followed stacked generalization ensemble [25] approach to ensemble four pre-trained CNN models. For that, as the preliminary step, four benchmarked CNN models i.e. VGG19, ResNET50, MobileNet and DenseNet are considered as base models for the proposed novel model. We have reused these pre-trained networks with the parameter they have

learned on ImageNet dataset. Also, we have removed their final softmax classification layer as it contains 1000 neurons. We have frozen all the intermediate layers of base models to keep their original trained weights. After that, the global average pooling is applied for dimensionality reduction and an output feature vector is created for each of the base model.

B. Spatial Feature Fusion of Deep Transfer Learning Features

Fusing the features obtained from various pre-trained CNN models supported many applications to achieve better performance than the conventional approach of utilizing the single deep CNN network for the task of classification [26]. The classification approach presented in this paper maps the feature space obtained by various pre-trained CNN models like VGG16, ResNet50, MobileNetV2 and DenseNet169 to a generalized feature space [26]. The feature space generated by each feature extractor with different dimension is represented as:

$$f_i = \{F_i^{(m)}\}_{m=1}^j \quad (1)$$

Where $i=1, 2, \dots, N$ represents the number of images in the dataset and $m=1, 2, \dots, j$ represents the number of pre-trained CNN models used as feature extractors. The cumulative feature set for a particular image sample is defined as:

$$f_k = \{f_k^1, f_k^2, f_k^3, \dots, f_k^l\} \quad (2)$$

Where, f_k^i is representation of i th dimension feature in the given k th sample and $i, i \in [1, D]$ represents the individual feature dimension. Thus, features obtained by the particular

CNN model forms individual feature space [27]. This, individual feature space is defined as:

$$\psi = \{f_1, f_2, \dots, f_n\} \quad (3)$$

Where $f_i, i \in [1, n]$ is the feature set of individual sample and n represents the total number of samples in the dataset.

Feature fusion technique combines heterogeneous features obtained from various CNN models and utilize combined features for comprehensive processing for the cumulative decision-making. The feature fusion technique combines feature spaces obtained from individual CNNs and provide feature subspace, which is generalized than the original feature space. The feature fusion technique to create a generalized feature space is represented as:

$$f: (\psi_p, \psi_q, \psi_r, \psi_s) \rightarrow \psi \quad (4)$$

Where, f is future fusion function and ψ_p, ψ_q, ψ_r and ψ_s represents the feature spaces obtained by individual CNN model. The data imbalance and noise are the major limitations of the features obtained using individual CNNs. However, obtaining generalize feature vector by combining different feature spaces helps in selecting important features captured by different CNN models which leads to more accurate classification accuracy.

C. Spatial Feature Fusion

In this research, the spatial feature fusion algorithm fuses the feature maps obtained by four different deep pre-trained CNN models [28]. Hence, four pre-trained CNN models are connected with each other using feature fusion techniques and point of connection between four models is known as fusion point. The training of softmax classifier is the next step after the fusion point in order to achieve the result of the classification as represented in Fig. 1. The spatial feature fusion function represented as:

$$f: K^p + K^q + K^r + K^s \rightarrow z \quad (5)$$

Where, K^p, K^q, K^r and K^s presents the set of features extracted by feature extractor P, Q, R and S respectively. Here, z denotes the fusion space features' set and $K^p, K^q, K^r, K^s, z \in \psi^{LWD}$ where L states the length, W states the width and D states the channels of the feature set, respectively [28].

D. Classification

After feature extraction, it is required to classify the data into various classes. This can be achieved using fully connected layers. In the proposed model, the third stage is performing classification. We have obtained a final feature vector after performing spatial feature fusion. For classification, two FC layers with ReLU and one FC layer with softmax are added to the proposed network as depicted in Fig. 1. A Rectified Linear Unit (ReLU) is used as an activation function for the first two fully connected layers. As the network classifies the input image into eight distinct classes, the last fully connected layer is applied with the softmax activation function.

IV. EXPERIMENTAL RESULTS AND DISCUSSION

The details regarding dataset, experimental setup and results are discussed in detail in this section. We have implemented the proposed ensemble CNN model using spatial feature fusion along with single model of VGG16, ResNet50, MobileNetV2 and DenseNet169.

A. Dataset

The experiment was performed on the microscopic images dataset that represents peripheral blood cell images [29]. The dataset contains 17902 microscopic jpg images with size of 360 X 363 pixels. The images are captured in the Core Laboratory at the Hospital Clinic of Barcelona. The images are taken using the analyzer CellaVision DM96. These all images are annotated by the clinical expert pathologists. The dataset contains very high-quality labelled images that can be used to train any algorithms belonging to machine learning and deep learning. The dataset categorized into eight distinct groups that are platelet, neutrophil, monocyte, lymphocyte, immature granulocytes, erythroblast, eosinophil and basophil. The dataset is specifically used for hematological diagnosis using computational tools. All the images of the dataset are acquired from the year 2015 to 2019 [30]. Fig. 2 represents the sample images from platelet, neutrophil and monocyte class.

The proportion of images in the dataset is represented in Fig. 3. As mentioned earlier there are 17902 microscopic images that categorized into 8 distinct classes.

From the dataset, 80% data are used for training and 20% data are used for validation.

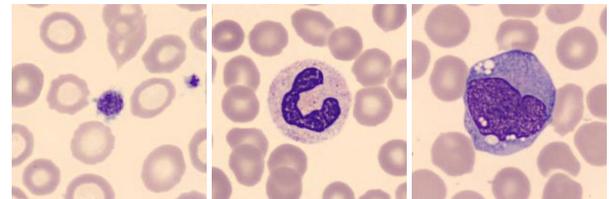


Fig. 2. Sample Images of Platelet, Neutrophil and Monocyte Classes.

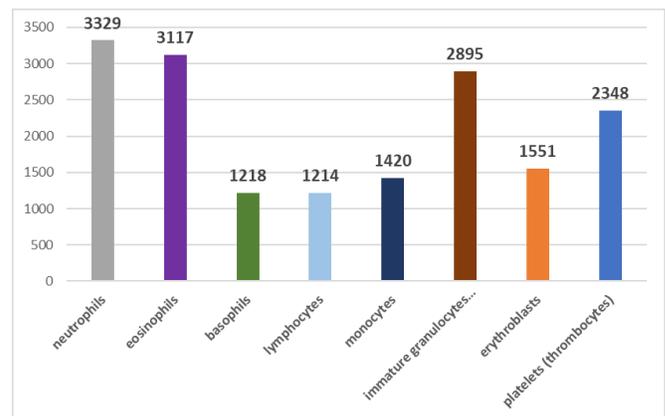


Fig. 3. Distribution of Microscopic Peripheral Blood Cell Dataset.

B. Environment Setup

The experiment is carried out on a workstation configured with an Intel® Core™ i7 8th generation processor, 32GB RAM, and NVIDIA Titan XP GPU with 64-bit Windows 10 operating system. To build the model for microscopic peripheral blood cell images classification, TensorFlow and Keras [31] API are used. TensorFlow is an open-source library and it provides many of the high-level and low-level APIs. Keras is one of the high-level API that built upon the TensorFlow library. It is a user-friendly and extensible library that makes the building of deep neural network fast. Keras offers various building blocks like activation functions, layers, optimizers, batch normalization etc. that are required to build a convolutional neural network. It provides the functionalities to run the CNN on a graphical processing unit (GPU) or tensor processing unit (TPU).

C. Network Training and Hyperparameter Tuning

One of the pre-requisite for any of the CNN model deployment is Training. One of the good practices is to use the pre-trained model instead of building and training the CNN model from scratch. Transfer learning is about re-using the pre-trained CNN model for solving a new problem [32]. All benchmark pre-trained CNN models are trained on the large dataset like ImageNet, PASCAL, COCO, etc. In this research, the weights along with learning of pre-trained model are transfer in the CNN model. Also, a fine-tuning process is performed by unfreezing the few of top layers to retrain them for classification purpose [33]. The dataset considered in the experiment is a small dataset as it contains 17902 images categorized into 8 distinct classes. Transfer learning and fine-tuning is performed by removing the last layer and introducing fully-connected softmax layer for the purpose of classification.

The basic building block of CNN model is artificial neural network. Thus, CNN models possess self-learning capability. For that, during training process, the weight of the network is adjusted in order to minimize the loss function [34]. There are set of parameters which controls the entire training process [35][36]. In this experiment, model specific hyperparameters are used. Table I summarizes the hyperparameters along with their optimized value.

TABLE I. OPTIMIZED VALUES FOR HYPERPARAMETERS

Hyperparameter (s)	Values
Number of Epochs	20
Learning rate	0.001
Batch size	32
Optimizer	SGD
Activation Function	ReLU / softmax
Dropout Rate	0.5

D. Evaluation Metrics

The benchmark evaluation metrics such as Precision, Recall, and F1-score are used to evaluate the performance of the proposed model. The equations used for the evaluation metrics are as per following.

$$\text{Precision}(r) = \frac{TP(r)}{TP(r)+FP(r)} \quad (6)$$

$$\text{Recall}(r) = \frac{TP(r)}{TP(r)+FN(r)} \quad (7)$$

$$F1 - \text{Score}(r) = \frac{2}{\frac{1}{\text{Recall}(r)} + \frac{1}{\text{Precision}(r)}} \quad (8)$$

Here, TP and FP are the true positive and true negatives, respectively. TN and FN are the true negatives and false negatives, respectively. To get the values of TP, TN, FP and FN, N X N confusion matrix is used, where N represents the number of classes available in the dataset [38]. The confusion matrix compares the actual values with the predicted values predicted by CNN model. It represents the values for TP, TN, FP and FN [37].

V. RESULT AND DISCUSSION

In this section, the performance of the proposed ensemble CNN model along with the performance of each individual CNN model i.e. VGG16, ResNet50, MobileNetV2 and DenseNet169 is discussed. To evaluate the performance of the model, two metric namely accuracy and loss are considered. The average accuracy and average loss for training and validation are measured. Accuracy is a metric used to measure the performance of a model. The accuracy is defined as:

$$\text{Accuracy} = \frac{TP+FN}{\text{Total no of samples}} \quad (9)$$

Where TP is the true positive and FN indicates false negatives [38].

Loss is defined as an error happened in prediction by a model. It is calculated by the difference between the value predicted by a CNN model and an actual value present in the dataset. While training the CNN model, the aim is to decrease the loss by optimizing the weights. Usually, two functions are used to measure a loss namely mean square error and cross-entropy. The cross-entropy is used to measure the loss and the formula is defined by the following equation [39].

$$\text{Cross - entropy} = - \sum_{i=1}^n \sum_{j=1}^n y_{i,j} \log(p_{i,j}) \quad (10)$$

Table II shows the values obtained for training and validation accuracy and loss respectively. It can be observed that the proposed CNN model achieved the highest accuracy i.e. 96.45% during training and 96.10% during validation. Moreover, it has the lowest loss i.e. 0.028 during training and 0.024 during validation. It shows that the proposed CNN model learns well as compare to the other models.

TABLE II. ACCURACY AND LOSS VALUES

Deep Learning Model	Training		Validation	
	Accuracy (%)	Loss	Accuracy (%)	Loss
Proposed CNN Model	96.45	0.028	96.10	0.024
DenseNet169	93.10	0.036	92.69	0.039
MobileNetV2	92.78	0.045	92.91	0.042
ResNet50	92.10	0.059	91.75	0.063
VGG16	87.56	0.097	86.80	0.089

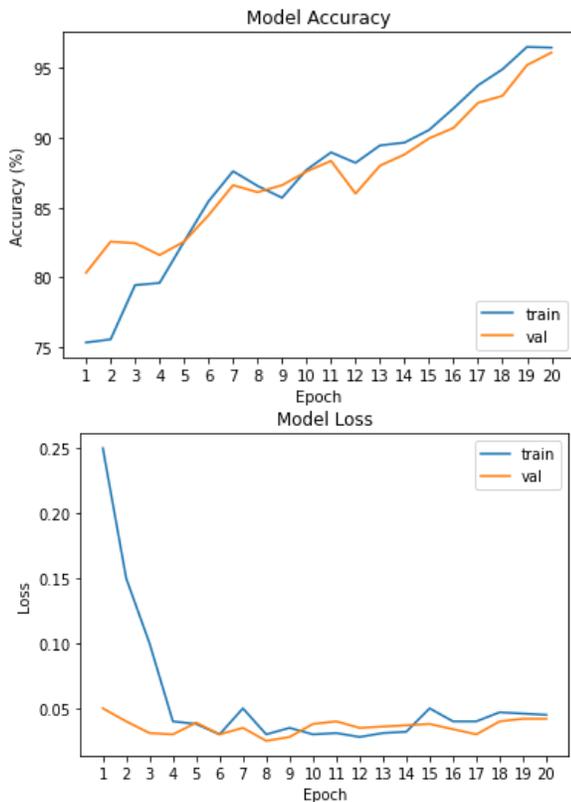


Fig. 4. Accuracy and Loss for the Proposed CNN Model.

The proposed CNN model also converges adequately and it can be shown from Fig. 4. The figure represents the accuracy and loss curves obtained during training and validation. Fig. 4 shows that the curves for both; training and validation for accuracy are going parallel and close to each other. Moreover, the loss curves are also going parallel and close to each other. This shows that model is trained adequately without facing the issues of underfitting and overfitting.

Table III represents the values for the precision, recall and F1-score for the different models obtained during experiment.

The F1-score, Precision and Recall values for VGG16 are 87.30%, 87.45% and 87.15 respectively. For ResNet50, the values are 92.17%, 92.25% and 92.10% respectively. MobileNetV2 provides 92.76%, 92.75% and 92.82% for F1-score, Precision and Recall. DenseNet169 gives 93.02%, 92.90% and 93.15% for F1-score, Precision and Recall. However, the proposed CNN model outperforms all other individual pre-trained CNN models and provides the highest values i.e. 96.55%, 96.40% and 96.70% for F1-score, Precision and Recall respectively.

From the results, it can be concluded that the spatial feature fusion approach for classification of microscopic blood cell images provides better results with enhancement in learning ability. The generalized transfer learning feature space developed by deep feature extractor reduces the problem of bias and variance while increase the performance. Hence, the proposed CNN model has better performance than the individual pre-trained CNN models.

TABLE III. PRECISION, RECALL AND F1-SCORE VALUES

Deep Learning Model	F1-Score (%)	Precision (%)	Recall (%)
Proposed CNN Model	96.55	96.40	96.70
DenseNet169	93.02	92.90	93.15
MobileNetV2	92.76	92.75	92.82
ResNet50	92.17	92.25	92.10
VGG16	87.30	87.45	87.15

VI. CONCLUSION

Deep learning models provide promising results in biomedical image analysis. It brings a new way for analysing and interpreting data that helps to identify the early signs of health conditions. A spatial feature fusion approach for biomedical image classification based on ensemble deep CNN and transfer learning is proposed in this paper. For that, the generalized transfer learning feature space is developed and a spatial fusion is applied to merge the learned features of different pre-trained CNNs. The paper covers the details of implementation and evaluation of the most proposed CNN model for classifying microscopic peripheral blood cell images. The dataset contains 17902 images of blood cell that are categorized into 8 class labels. The experiment shows that the proposed ensemble CNN model outperforms individual pre-trained CNN model and provides better precision, recall and F1-score values.

REFERENCES

- [1] "Biomedical imaging & image processing," *Embs.org*. [Online]. Available: <https://www.embs.org/about-biomedical-engineering/our-areas-of-research/biomedical-imaging-image-processing>. [Accessed: 19-Jan-2022].
- [2] "The important role medical imaging plays in diagnosis and treatment :: PBMC health," *Pbmchealth.org*. [Online]. Available: <https://www.pbmchealth.org/news-events/blog/important-role-medical-imaging-plays-diagnosis-and-treatment>. [Accessed: 19-Jan-2022].
- [3] U. F. O. Themes, "The role of imaging in medicine," *Radiology Key*, 09-Aug-2020. [Online]. Available: <https://radiologykey.com/the-role-of-imaging-in-medicine>. [Accessed: 19-Jan-2022].
- [4] Z. Lai and H. Deng, "Medical image classification based on deep features extracted by deep model and statistic feature fusion with multilayer perceptron," *Comput. Intell. Neurosci.*, vol. 2018, p. 2061516, 2018.
- [5] F. Pesapane, M. Codari, and F. Sardanelli, "Artificial intelligence in medical imaging: threat or opportunity? Radiologists again at the forefront of innovation in medicine," *Eur. Radiol. Exp*, vol. 2, no. 1, 2018.
- [6] I. Patel and S. Patel, "An optimized deep learning model for flower classification using NAS-FPN and Faster R-CNN," *Ijstr.org*. [Online]. Available: <http://www.ijstr.org/final-print/mar2020/An-Optimized-Deep-Learning-Model-For-Flower-Classification-Using-Nas-fpn-And-Faster-R-cnn.pdf>. [Accessed: 19-Jan-2022].
- [7] H.-H. Zhao and H. Liu, "Multiple classifiers fusion and CNN feature extraction for handwritten digits recognition," *Granul. Comput.*, vol. 5, no. 3, pp. 411–418, 2020.
- [8] M. Hassaballah, A. A. Abdelmgeid, and H. A. Alshazly, "Image features detection, description and matching," in *Image Feature Detectors and Descriptors*. Cham: Springer International Publishing, 2016, pp. 11–45.
- [9] S. Ge, C. Bai, Y. Liu, Y. Liu, and T. Zhao, "Deep and discriminative feature learning for fingerprint classification," in *2017 3rd IEEE International Conference on Computer and Communications (ICCC)*, 2017, pp. 1942–1946.
- [10] N. Sharma, A. K. Ray, S. Sharma, K. K. Shukla, S. Pradhan, and L. M. Aggarwal, "Segmentation and classification of medical images using

- texture-primitive features: Application of BAM-type artificial neural network,” *J. Med. Phys.*, vol. 33, no. 3, pp. 119–126, 2008.
- [11] A. M. Sayed, “Machine learning augmented breast tumors classification using magnetic resonance imaging histograms,” *Int. J. Adv. Comput. Sci. Appl.*, vol. 12, no. 12, 2021.
- [12] M. I. Daoud, T. M. Bdair, M. Al-Najar, and R. Alazrai, “A fusion-based approach for breast ultrasound image classification using multiple-ROI texture and morphological analyses,” *Comput. Math. Methods Med.*, vol. 2016, p. 6740956, 2016.
- [13] P. Chak, P. Navadiya, B. Parikh, and K. C. Pathak, “Neural network and SVM based kidney stone based medical image classification,” in *Communications in Computer and Information Science*, Singapore: Springer Singapore, 2020, pp. 158–173.
- [14] P. Nanglia, S. Kumar, A. N. Mahajan, P. Singh, and D. Rathee, “A hybrid algorithm for lung cancer classification using SVM and Neural Networks,” *ICT Express*, vol. 7, no. 3, pp. 335–341, 2021.
- [15] H. T. Nguyen, T. Bao, H. Hoang, T. Phuoc, and Nghi, “Improving disease prediction using shallow convolutional neural networks on metagenomic data visualizations based on mean-shift clustering algorithm,” *Int. J. Adv. Comput. Sci. Appl.*, vol. 11, no. 6, 2020.
- [16] S. M. Anwar, M. Majid, A. Qayyum, M. Awais, M. Alnowami, and M. K. Khan, “Medical image analysis using convolutional neural networks: A review,” *J. Med. Syst.*, vol. 42, no. 11, p. 226, 2018.
- [17] C.-H. Chiang, C.-L. Weng, and H.-W. Chiu, “Automatic classification of medical image modality and anatomical location using convolutional neural network,” *PLoS One*, vol. 16, no. 6, p. e0253205, 2021.
- [18] B. P. Battula and D. Balaganesh, “Medical image data classification using deep learning based hybrid model with CNN and encoder,” *Rev. d intell. artif.*, vol. 34, no. 5, pp. 645–652, 2020.
- [19] A. A. Goma and Y. M. A. El-Latif, “Early Prediction of Plant Diseases using CNN and GANs,” *Int. J. Adv. Comput. Sci. Appl.*, vol. 12, no. 5, 2021.
- [20] S. Patel, “Bacterial colony classification using atrous convolution with transfer learning,” *Annals of RSCB*, pp. 1428–1441, 2021.
- [21] M. A. Ganaie, M. Hu, A. K. Malik, M. Tanveer, and P. N. Suganthan, “Ensemble deep learning: A review,” *arXiv [cs.LG]*, 2021.
- [22] S. Patel and R. Patel, “A comprehensive analysis of A Comprehensive Study of Applying Convolutional Neural Network for Computer Vision,” *Int. j. adv. sci. technol.*, vol. 29, no. 6s, pp. 2161–2174, 2020.
- [23] O. Russakovsky *et al.*, “ImageNet large scale visual recognition challenge,” *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, 2015.
- [24] N. Ganatra and A. Patel, “Deep learning methods and applications for precision agriculture,” in *Machine Learning for Predictive Analysis*, Singapore: Springer Singapore, 2021, pp. 515–527.
- [25] J. Brownlee, “Stacking ensemble for deep learning neural networks in python,” *Machine Learning Mastery*, 30-Dec-2018. [Online]. Available: <https://machinelearningmastery.com/stacking-ensemble-for-deep-learning-neural-networks/>. [Accessed: 21-Mar-2022].
- [26] T. Akilan, Q. M. J. Wu, Y. Yang, and A. Safaei, “Fusion of transfer learning features and its application in image classification,” in *2017 IEEE 30th Canadian Conference on Electrical and Computer Engineering (CCECE)*, 2017, pp. 1–5.
- [27] Y. Lavinia, H. H. Vo, and A. Verma, “Fusion based deep CNN for improved large-scale image action recognition,” in *2016 IEEE International Symposium on Multimedia (ISM)*, 2016, pp. 609–614.
- [28] S. Zhou, S. Cai, C. Zeng, and Z. Wang, “A cotton and flax fiber classification model based on transfer learning and spatial fusion of deep features,” in *Advances on P2P, Parallel, Grid, Cloud and Internet Computing*, Cham: Springer International Publishing, 2020, pp. 152–162.
- [29] A. Acevedo, S. Alférez, A. Merino, L. Puigví, and J. Rodellar, “Recognition of peripheral blood cell images using convolutional neural networks,” *Comput. Methods Programs Biomed.*, vol. 180, no. 105020, p. 105020, 2019.
- [30] A. Acevedo, A. Merino, S. Alférez, Á. Molina, L. Boldú, and J. Rodellar, “A dataset of microscopic peripheral blood cell images for development of automatic recognition systems,” *Data Brief*, vol. 30, no. 105474, p. 105474, 2020.
- [31] Keras Team, “Keras: the Python deep learning API,” *Keras.io*. [Online]. Available: <https://keras.io>. [Accessed: 19-Jan-2022].
- [32] A. P. Ganatra, “Performance analysis of fine-tuned convolutional neural network models for plant disease classification,” *Int. J. Contr. Autom.*, vol. 13, no. 03, pp. 293–305, 2020.
- [33] J. Browarczyk, A. Kurowski, and B. Kostek, “Analyzing the effectiveness of the brain-computer interface for task discerning based on machine learning,” *Sensors (Basel)*, vol. 20, no. 8, p. 2403, 2020.
- [34] A. Brodzicki, J. Jaworek-Korjakowska, P. Kleczek, M. Garland, and M. Bogyo, “Pre-trained deep convolutional neural network for *Clostridioides difficile* bacteria cytotoxicity classification based on fluorescence images,” *Sensors (Basel)*, vol. 20, no. 23, p. 6713, 2020.
- [35] K. Nongthombam and D. Sharma, “Data Analysis using Python,” *Int. J. Eng. Res. Technol. (Ahmedabad)*, vol. 10, no. 7.
- [36] R. Susmaga, “Confusion Matrix Visualization,” in *Intelligent Information Processing and Web Mining*, Berlin, Heidelberg: Springer Berlin Heidelberg, 2004, pp. 107–116.
- [37] L. Cavallaro, O. Bagdasar, P. De Meo, G. Fiumara, and A. Liotta, “Artificial neural networks training acceleration through network science strategies,” *Soft Comput.*, vol. 24, no. 23, pp. 17787–17795, 2020.
- [38] I. Patel and S. Patel, “Flower Identification and Classification using Computer Vision and Machine Learning Techniques”, *International Journal of Engineering and Advanced Technology (IJEAT)*, vol. 8, no. 6, pp. 277–285, 2019.
- [39] E. Gordon-Rodríguez, G. Loaiza-Ganem, G. Pleiss, and J. P. Cunningham, “Uses and abuses of the cross-entropy loss: Case studies in modern deep learning,” *arXiv [stat.ML]*, pp. 1–10, 2020.

Effectivity Score of Simulation Tools towards Modelling Design in Internet-of-Things

Gauri Sameer Rapate
Assistant Professor
Dept. of Computer Science and Engineering
PES University, Bangalore, India

Dr. N C Naveen
Prof and Head
Dept. of Computer Science and Engineering
JSS Academy of Technical Education, Bangalore, India

Abstract—Simulation tools play an integral and significant role in studying the applicability and effectiveness of different algorithms for solving real-world problems cost-effectively. In the case of Internet-of-Things, the issues associated with real-world implementation are exponentially multi-fold. Although various simulators have facilitated the evolution of schemes to address the problems in IoT applications in the last few years, their applicability in the real world is highly questionable. Hence, this paper discusses the potential features of existing simulations (both commercial and research-based) and investigates features of different assessment environment tools to understand their current state. The paper further contributes toward a distinct research pattern. The core contribution of this manuscript is to review standard practices of using simulation tool along with different test environments. The paper also contributes towards exploring various prospects of unaddressed problems associated with a usage of existing simulation environment/tool for investigating the challenging and practical environment of an IoT ecosystem. The learning outcome of this study will assist the reader to make a decision towards adopting precise simulation tool for their work as well as it also highlights the need to perform more number of customization towards including the features that is found in research gap.

Keywords—Internet-of-things; real-world; application; simulation model; environment

I. INTRODUCTION

The process and various operation ranges involved in the real world are sometimes quite challenging to investigate and study, considering real-world entities. This problem can be solved by using simulation modeling to efficiently and safely perform the processes to offer solutions [1]. The implemented methods using different spectrums of the algorithm during the simulation study facilitate a simplified verification process over-controlled research environment [2]. This paper discusses the simulation perspective of studying various problems in Internet-of-Things (IoT), a network of different physical devices called things with distinct identifiers [3]. The dependable attributes for framing an IoT ecosystem are communication devices, sensors, processors, etc. [4]. An edge device or an IoT gateway node collects the sensory data shared via IoT devices and is forwarded to the cloud for storage or analysis [5]. However, implementing such a large ecosystem of an IoT is comprehensively a massive task as they are also associated with various implementation impediments, e.g., challenges of inappropriately aggregated data, processing of

unstructured data, ensuring coverage and connectivity of a large number of IoT device, ensuring power management for resource constraint sensors, challenges associated with data storage, integrating with all points of an IoT with appropriate software, proper identification of an IoT device with legitimate authentication, and compatibility issues of a large number of heterogeneous system in an IoT [6]. However, it should be noted that IoT is not only about sensory data collection [7], routing [8], and storage [9], but it is strongly related to potential data analytical operations, too [10]. At present, there are various schemes to ensure an effective operation of IoT communication in the majority of the perspective [11]-[14]; however, it is seen that the majority of such studies are carried out using a simulation-based approach while only a few proportions of work is carried out over real-time models and data. Both such schemes (real world and simulation-based) have their benefits and limitation. However, simulation-based studies are still more favorable toward cost-effective analysis, leading to much successful deployment in the real world. However, such a simulation-based approach should be deeply investigated to find out if they are really at par with solving impending problems in an IoT.

There is no doubt about the availability of a list of different simulators in current times for deploying novel logic for implementing an IoT. However, questions still arise about the taxonomies of such simulators, the discrete beneficial features towards cost-effective deployment, judging the applicability of simulation outcomes in the practical world, and finding the most adopted simulation tool in recent years. Finding answers to these questions will eventually guide better decision-making, either towards adoption or the evolution of new simulators in IoT. The prime motivation of the proposed study is to showcase the discussion on simulators with a higher and lower adoption rate in the last few years. Therefore, the *novelty* of the proposed study is towards assessing the functionalities and features of existing simulators towards investigating problems in IoT. Finally, the study also elaborates on the first and second levels of the research gap, unlike any existing research update. The contribution of the paper is as follows:

1) The paper offers a brief and compact insight towards the existing 5 standard simulation tools of an IoT exercised in commercial sector as well as 16 frequently adopted simulation tools practiced in research-based area.

2) The paper highlights essential characteristics of 6 standard environments adopted for assessing the protocols and algorithms meant for an IoT environment.

3) Existing trends of research work that is found towards developing simulation-environment is discussed in this paper that assists in highlighting the highly adopted simulation tool for studying IoT ecosystem.

4) The paper exclusively discussed elaborately about the research gap associated with the existing developments towards IoT simulation tool that offers critical information about unaddressed research problems.

The manuscript's organization is as follows: Section II discusses the existing simulation tools into practice, while Section III discusses the assessment environment offered by such simulation tools. Discussion of existing research trends towards using simulation tools is carried out in Section IV. In contrast, a meeting of the research gap is carried out in Section V. Finally; Section VI provides conclusive remarks and future work direction.

II. SIMULATION TOOLS FOR IOT

The simulation tools for assessing the performance of a newly developed logic for an IoT environment must adhere to standard vital parameters. With an expected budget of communication link above 164 dB and modulation method of either BPSK or QPSK for uplink transmission and QPSK for downlink transmission, the simulation for an IoT should be set up in FDD half-duplex mode. The simulation key parameter includes 64 kbps and 25 kbps of uplink and downlink data transmission with a latency of around 10 seconds. The most simulation also considers SC-FDMA for uplink transmission while OFDM for downlink transmission. The above mentioned are the prevalent values of critical parameters to be considered while carrying out a simulation study. This section further highlights the frequently used standard simulation tool in an IoT as follows:

A. Commercial Simulation Tool

These are the commercially used tools to assess the conditional logic, novel logic for communication in an IoT, and during prototyping of certain sense of implementation in an IoT.

1) *MIMIC simulation tool [15]*: This simulator manages various essential entities in an IoT environment, e.g., all the connected devices, sensors, and gateway nodes. The idea is to form a standard test assessment scenario of IoT capable of deploying and evaluating Industry 4.0, architectures with event-driven approaches, agriculture, factories, and smart cities.

2) *IoTNetSim tool [16]*: This simulation tool can manage different variants of the IoT network environment and structural information of the heterogeneous IoT nodes. The core idea is to facilitate a better form of modeling toward network connectivity concerning target application logic. The complete simulation process is carried out over three layers, viz. i) cloud layer, ii) edge layer, and iii) IoT layer. The *cloud*

layer is responsible for obtaining and managing the data processed by the edge layer, particularly in the data center of an IoT. The processed information is then forwarded to the sophisticated virtual machine via a specific set of the authenticated host. The *edge layer* is responsible for obtaining the data from the gateway node, followed by further processing the data to forward it to the devices that run under the supportability of the cloud layer. The *IoT layer* is responsible for deploying and generating the sensing devices to transmit the data to the link node. Further, this data is forwarded to the gateway node that can access and process this data to be forwarded to the edge layer in the form of aggregated information.

3) *Cooja simulation tool [17]*: This is a frequently used simulation tool mainly for sensory application. Due to this capability, it is also used for an IoT environment with the involvement of sensors. This simulation tool can evaluate sensors with intelligent capabilities, different communication technologies, and frequently used internet protocols in an IoT. This simulation tool uses Contiki mote to simulate the real-time assessment environment capabilities. Three forms of windows carry out the complete operation of this simulation tool, i.e., i) network window, ii) control window for simulation, and iii) note window. The *network window* can exhibit the environments of all mores and their respective radio traffic. It can also organize the physical elements associated with the sensor mote for a more in-depth and practical analysis of sensory devices in IoT. The *control window for simulation* is responsible for managing all the execution during simulation and simulation speed. Finally, the *note window* can store the execution logic and contains all the essential simulation points. It should be noted that the standard protocol of 6LoWPAN is adopted by this simulation tool. Apart from this, the standard protocol of IPv6 is also adopted by this tool concerning devices compliant with the family of IEEE 802.15.4 standards. This prime dependency of this simulation tool to develop an IoT environment are mainly two viz. i) Contiki operating system as a part of a software module and Tmote Sky as a part of the hardware module. Using both software and hardware module, this simulation tool can perform initialization of sync nodes and sensor nodes followed by establishing a reliable communication using standard protocols. Further, the sensor node is transmitted to the sink node from the transmitting node by this tool.

4) *IBM Bluemix [18]*: This is one of the famous commercial simulation tools for assessing the prototypes without dependencies on the physical device associated with an IoT platform. It is an IoT platform made to be functional in the IBM open cloud system so that the developer can easily access the proprietary software of IBM. This can significantly evaluate certain essential vital functions and security features associated with mobile and web applications. The continuous availability of cloud services offers the practical application and service manageability.

5) *SimpleIoTSimulator* [19]: This is one of the simplified forms of commercial simulator used to assess the environment for many IoT devices. It also offers much supportability of conventional IoT protocols, e.g., HTTPS, CoAP, and MQTT. This simulation tool facilitates different vendors to enhance product quality by formulating a better test environment. It offers supportability of both IPv6 and IPv4 sensors with operational supportability towards the constraint environment of an IoT.

B. Research-based Simulation Tool

There is a big difference between the simulation tool adoption for commercial practices and research-based studies. The commercial rules call for simulation tools with specific evaluation features; however, this is not the case with research-based simulation studies. Research work towards an IoT environment calls for including many features, conditional logic, and a flexible deployment environment to testify the targeted logic. Hence, simulation tools adopted for research-based work offer a more discrete set of operations with comprehensive functionalities. This section highlights the existing research-based simulation tools (Table I) as follows:

1) *ANSYS-IoT simulator* [20]: This form of simulation tool analyzes various forms of challenges in an IoT. The core usage of this simulation tool is mainly to assess the cumulative integrity of sensing devices with innovative capabilities deployed in an IoT environment. It is also used to assess longevity and energy consumption and enhance reliability, followed by validation.

2) *Bevywise simulation tool* [21]: This tool is specifically used for the scenario when fog computing is used in collaboration with an IoT environment. The tool facilitates the virtual clients by deploying the MQTT protocol on-premise. It also offers different variants of functional assessment over the cloud environment by enabling the deployment of many commodity servers. Particularly helpful for the industrial IoT environment, this simulation tool offers an end-to-end solution considering the constraints of the practical world of IoT implementation.

3) *IoTIFY* [22]: This tool offers a software backbone for operations associated with hardware modules of an IoT. This tool provides a precise building of a digital lab by harnessing a simplified construction of an analytical model and virtualization of an IoT device. It can be used for analytical scaling, solutions for load testing, and building prototypes of an embedded system. It also assists in generating records and can well manage IoT devices used in a fog computing environment.

4) *EdgeCloudSim* [23]: This simulation tool caters to the demands of IoT operation dependent on computational queries of edge devices. It can carry out sophisticated calculations and evaluate its capacity to process the query and allocate necessary resources. This simulation tool is mainly meant for developers to assess the impact of their machine parameters on the communication and processing needs of an IoT node.

5) *MobIoTSim* [24]: The majority of the existing simulators of IoT demand higher resources to be used and are meant to be run on computing devices, e.g., desktops or laptops. This simulation tool is intended to simulate mobile-device built on the Android operating system.

6) *TOSSIM* [25]: This is another frequently used simulator by researchers who adopts sensor mote running on TinyOS. Hence, this tool can assess the smart sensory application running over an action-based operating environment.

7) *SWE Simulator* [26]: This simulation tool is mainly meant to assess the large-scale deployment scenario of IoT, including sensors. The tool also assists in effective cost control during implementation scenarios. It can also evaluate the presence of any risk factor and its possible influence. The simulation tool also uses observation services of sensors to develop topology and assess the performance of the assessment in an IoT.

8) *Atomiton* [27]: This is another simulator capable of simulating all the innovative sensory applications and involving different forms of actuators. Hence, this simulation tool assesses a specific state of the operating environment used for the complex operation of smart sensors in an IoT.

9) *QualNet* [28]: When the simulation study demands high-end accuracy, this is a preferred simulation tool for IoT applications. However, it should be noted that not all sensor devices are used for this form of simulation. Only the sensors compliant with specific families of IEEE 802.15.4 consider smart sensors. Although this is a commercial simulation tool, they are used extensively for research-based study. It also offers enriching interactivity owing to the improvised form of the Glomosim simulator.

10) *NS3 Simulator* [29]: This frequently used network simulator considers networks of the practical world scenario in IoT. The prime usage of this simulator is to assess the robustness of security features to resist various threats in the communication environment of an IoT. The tool offers the developer to use python and C++ for scripting while making use of the standard protocol of LTE, ZigBee, and 6LOWPAN. While working in an IoT environment, this tool adds three discrete types of devices, e.g., IoT node, blockchain, and Gateway. This simulation toolbox also assists in routing the class of all these entities in an IoT. The prime set of operations in an NS3 is to deploy the sensing IoT nodes, followed by applying standard protocol implementation of MQTT, HTTP, and CoAP. Finally, the data is forwarded to the sink.

11) *OMNet++* [30]: This simulation tool is popularly known for its discrete event simulation facilitation. A unique discrete simulation environment is constructed using enriched libraries from C++. Inspired by the Eclipse environment, this tool uses high-level language to develop more prominent components. It thereby forms an object-oriented model of simulation that could be used for different apps.

12) *SimIoT* [31]: This simulation tool can be used for any IoT system characterized by the static or dynamic attributes in a trial and error process. This mechanism is primarily meant

for managing the clouds environment system concerning its host of data centers. It assists in both method construction and the configuration of various entities.

13) *CupCarbon* [32]: This simulation tool is mainly used for assessing wireless sensor networks in their discrete form using multi-dimensional visualization features with multi-agents. Different forms of distributed algorithms can be validated using this simulation tool. It can also validate various iterative tasks over automation processes mainly meant for industrial applications. Further, it can also evaluate different forms of routing schemes, protocols, communication, and topologies in wireless networks involved in an IoT.

14) *IoTSim* [33]: This is another popular simulation tool used in IoT deployed alongside cloud environments. It can also be an extension of the popular cloud-based simulation CloudSim. One of the significant functions of this simulation tool is that it assists in processing big data that uses distributed software framework MapReduce. The process of identification,

as well as outcome analysis, is carried out using this simulation tool. It can be used for both research-based studies and industrial evaluation.

15) *iFogSim* [34]: This simulation is meant to carry out multiple levels of evaluation-based operation associated with different environment variants. It can simulate network connection, edge devices, fog data centers, IoT, etc. It can also be used for assessing another performance metric to be carried out over an IoT cloud environment. This simulation tool can also perform extensive assessments of various QoS metrics.

16) *DPWSim* [35]: This simulation tool profiles the devices synced with the web services associated with information exchanges, classification of services, and effective identification. This simulation tool is mainly used for explicit devices and applications related to the IoT to assess the dual form of operation, i.e., a function that enables hosting and processes that are being hosted. It also offers a better form of secure exchange of data

TABLE I. SUMMARY OF RESEARCH-BASED SIMULATION TOOL

<i>Tool</i>	<i>Security Score</i>	<i>Service domain</i>	<i>API integration</i>	<i>Built-in IoT Standards</i>	<i>A layer of IoT Architecture</i>	<i>Programming</i>	<i>scope</i>	<i>Type</i>
Ansys-IoT	High	Industry	REST	Real-Time	Network	Java, Python	IoT industry	Autonomous
Bevywise-IoT	Medium	Smart City	REST	Real-Time	Network	Java, Python	IoT device	Broker
IoTIFY	High	Smart City	REST	Real-Time	Application, Network	Java, Python	Hardware connection	Mobile App
Edge CloudSim	High	Edge Orchestrator	SOAP	Mist Computing	Network	Matlab	Edge, WLAN	Realistic
MobIoTSim	Medium	Generic	REST	Device profile for web services	Application, network	C#, C++	IoT Network	Research-based
TOSSIM	High	Generic	REST	Injecting packets	Communication Network	Python, C	TinyOS	Sensor monitoring
SWE-IoT	High	Human Interface	SOAP	Collision Detection	Communication Network	C, C++	WSN	Sensor monitoring
Atominition IoT	High	MQIdentity	REST	Socialize	Communication Network	Java	IoT, IIoT	Edge
QualNet	Medium	Generic	REST	802.15.4	Perceptual Network	C++	Network	Discretization
NS3	High	Generic	REST	LoRaWAN	Perceptual Network	C++	Network	Discrete event
OMNeT++	Medium	Generic	SOAP	Manual extension	Perceptual Network	C++	Network	Discrete event
SimIoT	High	Generic	REST	No	Application	Java	Data Analysis	Discrete event
CupCarbon	High	Smart city	UDX	LoRaWAN, 802.15.4	Perceptual Network	Java	Network	Discrete event (agent)
IoTSim	Medium	Generic	REST	No	Application	Java	Data Analysis	MapReduce model
iFogSim	Medium	Generic	SOAP	No	Perceptual Network Application	Java	Fig	Discrete event
DPWSim	Medium	Generic	SOAP	Messaging Web services security	Application	Java	IoT	Open-source

III. ASSESSMENT ENVIRONMENT OF IOT

The previous section has elaborated on various existing simulation tools for evaluating various problems and their respective solution. It should be noted that the simulation tool facilitates the user to deploy their logic of implementation by reducing the same problem space of the IoT environment. It also enables various libraries to develop novel solutions. However, the simulation tool itself cannot be assumed to be 100% fulfilling the outcomes of the simulation study. For this purpose, there is a need for a legal assessment environment to be considered while carrying out the simulation study in IoT. Hence, the assessment environment provides a spectrum of IoT environments to analyze the solution model of researchers and assess troubleshoots, debugging, developing, and creating new logic, which is universally accepted. The prime advantage of considering a legal assessment environment is that it offers practical world usage of devices, interactions with the operating system, remote administration, and the capability to execute real-world devices/services/applications. Some of the standard assessment environments for simulating an IoT are JOSE [36], Smart Santander [37], FIESTA-IoT [38], FIT IoT-LAB [39], WHYNET [40], and MBTAAS [41] that are briefed as follows:

1) *JOSE [36]*: This assessment environment is meant for evaluating the devices or services associated with outdoor communication. It also offers many subject trials to facilitate resource computation, sensor network deployment, storage management, etc.

2) *Smart Santander [37]*: This is another standard assessment environment in an IoT that facilitates the adoption of many IoT devices along with small radio-based services and identification code deployment over both the static and mobility aspects of the nodes. This tool can evaluate traffic intensity mainly for the mobile environment in an IoT.

3) *FIESTA-IoT [38]*: This is one of the giant assessment environments considered during IoT simulation to offer its semantic and interoperable assessment feature. It integrates multiple numbers of another assessment environment in an IoT

to analyze the corpus of data. This large data further assists in facilitating webservices in live stream mode. The highly interconnected systems in this tool offer a significant ability to exchange information among various federated assessment environments in an IoT.

4) *FIT IoT-LAB [39]*: This form of assessment environment is made for performing experimentation on IoT on a huge scale. Various objects can be broadcasted and developed by this tool, considering many low-resource nodes for assessment. It also uses mobile robots to testify various upcoming innovative applications in IoT.

5) *WHYNET [40]*: Essentially meant for hybrid networks, this assessment tool is for performing realization of applications, protocols of WSN, heterogeneous networks, etc. Different forms of emulation and physical entities are carried out by this assessment tool using its single end interface itself. This environment can assess adaptive networks and large/small/medium scale networks in an IoT.

6) *MBTAAS [41]*: This tool is meant to offer assessment in the form of a service model. It also provides first-hand experience working over an IoT and getting acquainted with its functionality. The tool also provides various test-case formulations and solutions to multiple services to carry out an on-premise assessment of cloud-based IoT applications and services.

Table II summarizes the characteristics of the existing assessment environment of an IoT. One of the potential beneficial factors of using a standard assessment environment is that it enables all the entities and devices to carry out interactivity during testing that device. Hence, without constructing an actual physical device, adopting a standard assessment environment assists in truly justifying the effectiveness of its operation when exposed to a real-time environment. Although it is strongly advisable to use such a traditional assessment environment while carrying out a simulation study for an IoT, there are still various challenges. The primary difficulties are higher dependencies on assessment area, mobility of hubs, scaling operation, and repeatability.

TABLE II. SUMMARY OF EXISTING ASSESSMENT ENVIRONMENT

<i>Tool</i>	<i>Virtualization Support</i>	<i>Service domain</i>	<i>API integration</i>	<i>Built-in IoT Standards</i>	<i>A layer of IoT Architecture</i>	<i>Programming</i>	<i>scope</i>	<i>Type</i>
JOSE [36]	Distributed cloud	Real-time	SOAP	Sensor network	Virtualized Network	Javascript, Java, C	SDN, WSN, IoT	Smart ICT platform
Smart Santander [37]	Management console	Smart city	REST	802.15.4 RFID	Application Network	JavaScript, Java	Mobile sensing	Map data
FIESTA-IoT [38]	Meta Cloud	Ambient Environment	REST	Energy consumption	Communication Network	Python, Java, C	Energy	Sensor Monitoring
FIT-IoT LAB [39]	FIT Cloud	Heterogeneous platform	REST	802.15.4 LoRaWAN	Perceptual Network	Java, nesC	IoT Network	IoT spectrum
WHYNET [40]	Web Portal	Energy	SOAP	Application	Network	Java	Wireless	Network Protocol
MBTAAS [41]	IoT dashboard	Smart city	REST	Model-based	All	OCL	IoT Platform	Service Oriented

IV. EXISTING RESEARCH TRENDS

To understand the existing research trends, manuscripts published in the last six years have referred to the explicit usage of two different variants of simulation tools.

A. Adoption of Commercial Simulation Tools

There is a total of 256 manuscripts in IEEE Xplore digital library which has reportedly used commercial simulation tools, with 243 conference papers and 13 journal publication (Fig. 1).

They are mainly used to investigate problems associated with traffic management, object monitoring, indoor agriculture, security analysis, the discovery of resources, etc. Out of all this, some of the significant literature has been witnessed on IoTNetSim [16], Cooja [42]-[52], and IBM Bluemix [53]-[54] only. No significant modeling is being carried out towards using the MIMIC simulation tool and SimpleIoTSimulator. The advantage explored towards such adoption of commercial tools found are i) it offers a point-to-point exploration process for the target problem, ii) various prototyping using hardware are feasible to be investigated, iii) specific product or service-based analysis can be easily carried out. While the limitation found toward such adoption are i) it requires explicit skill to work on such tools, ii) various add-ons and software patches are required to be acquired to experience the full-fledged operation, iii) it doesn't offer extensible, cross-platform libraries for heterogeneous products/applications/services in IoT.

B. Adoption of Research-based Simulation Tools

One hundred seventy manuscripts are being reported to adopt research-based simulation tools, consisting of 126 conference papers and 44 journal papers (Fig. 2).

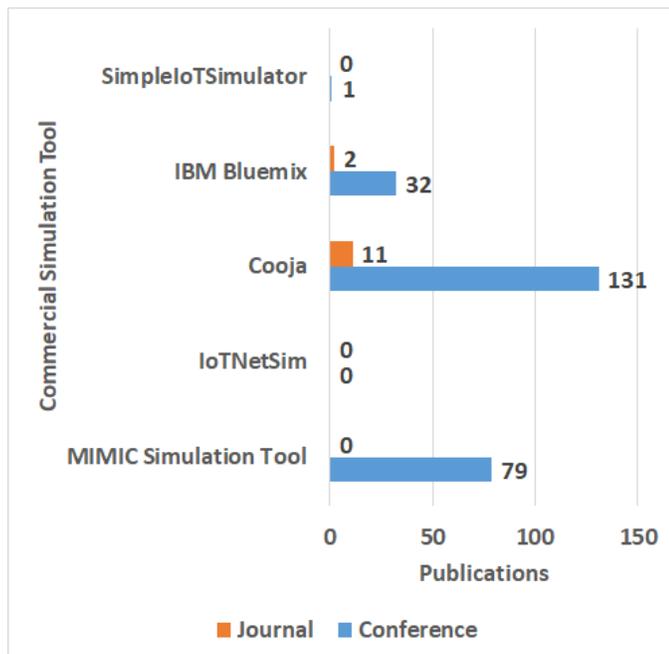


Fig. 1. Trends for Commercial on Simulation Tools.

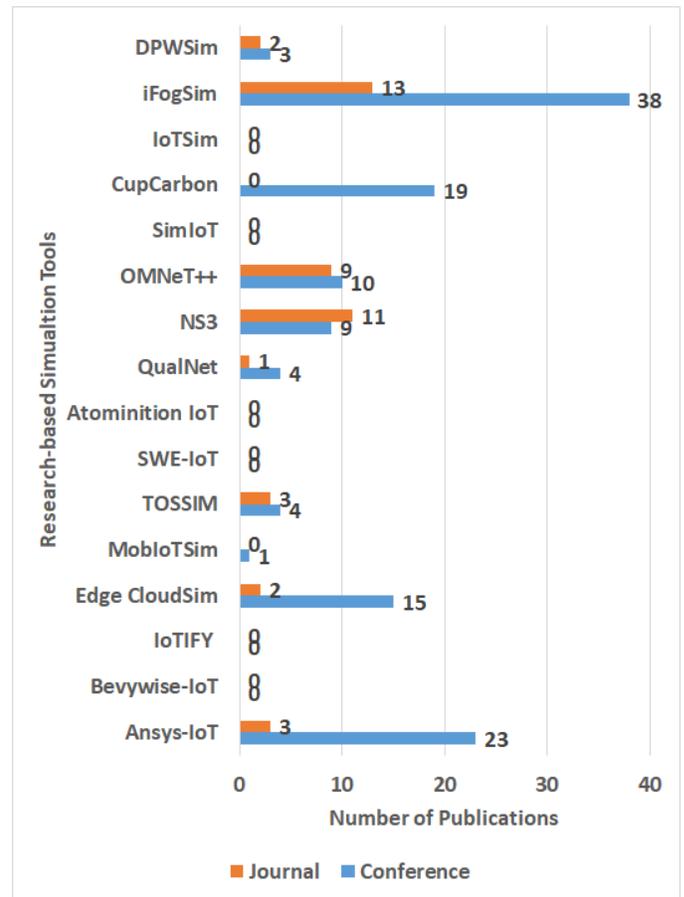


Fig. 2. Trends for Research-based Simulation Tools.

This will eventually mean that the adoption of research-based simulation tools is a bit lesser than commercial simulation tools. It is also noted that the adoption of each tool is used to solve some common problems and specific problems. The common problems will include traffic monitoring and security analysis, while the specific problems investigated by these tools will be scheduling, data transmission, application-specific evaluation, etc. Some of the significant literatures that has reported towards using Ansys-IoT [55]-[57], Edge CloudSim [58][59], TOSSIM [60]-[62], QualNet [63], NS3 [64]-[74], OMNeT++ [75]-[83], iFogSim [84]-[96], and DPWSim [97][98]. The beneficial points of this adoption are: i) the majority is open-sourced, ii) user-friendly, iii) extensible environment for analysis, and iv) it doesn't require complex configuration or setup. While the limiting factors are: i) inbuilt methods and libraries are sometimes challenging to match with the problem space of analysis, ii) each simulator has distinct features and functionalities while migration or integration is sometimes not possible, and iii) quite specific to environmental usage in IoT.

V. LITERATURE REVIEW

From the prior section, it is noted that not all the simulation tools are widely used either in case of commercial or research-based practices. Cooja is highest used in commercial simulation while iFogSim is majorly used for research-based study. It should be noted that all simulators, in either of the

categories, have some common functionalities as well as some exclusive functionalities. The prime contribution of the manuscript is actually to judge the constraints as well as limiting factors associated with existing simulation tools. This is the core reason that the biometric analysis of this paper mainly consists of only studies that adopted 103 sources which has adopted simulation tools. The analysis of some of the recent work towards investigation IoT has also developed a computational framework on the basis of customized simulation study [99]-[103]. However, it is to be noted that adoption of simulation is one of the critical decision to be made by the researcher on the basis of the problems to be addressed in their model. An effective simulation tool should offer higher flexibility to assess the algorithm or protocol without much skill and re-engineering process. Unfortunately, this is not the case with existing studies as researchers tends to adopts frequently used standard simulation tools, which can cater up their investigation objectives. This is done by overlooking next sequence of investigation to be carried out in line of current research work. Apart from this, majority of existing simulation tool discussed in this paper doesn't offer much of customization privilege. Therefore, after reviewing the complete papers, this section discusses about the open-end issues in the form of contribution and thereby these findings are novel compared to any existing review work being carried out in current times.

A. Discussion of Research Gap

After reviewing the existing features of two variants of the simulation tool, standard assessment environment during simulation study, and research pattern, it is found that there are various contributory factors as well as open-end research issues. This section explicitly highlights the research gap in the first and second levels of research gap for better understanding.

1) *First level of investigation:* The first level of research gap will eventually mean all the open-end issues retrieved during the reviewing problem. IoT, which consists of highly interconnected devices in large numbers, consists of a multidisciplinary domain with multiple challenges that are quite difficult to assess. The existing simulation tools offer good privilege for investigating semantic queries, protocols, data transmission, routing, usability, privacy, and security considering various applications or services without considering scalability issues. At the moment of transition of problems associated with any factor from intra-net to inter-net, there is eventual evolution of scalability issues. Either commercial or research-based simulation tools cannot detect this. Although it is a well-known fact that repeatability of analysis is quite challenging in IoT, after knowing this fact, existing simulators don't offer much privilege toward addressing increasing heterogeneity of devices and information associated with it. Considering any use-cases of an IoT to be testified over current simulators, it is inevitable to assess multiple application domains (e.g., smart city, vehicular IoT, Industrial IoT, etc.). It will mean that such inclusion will extensively maximize the concurrency towards accessing the infrastructure. This is a computationally challenging task that is

not facilitated by any existing simulators, and developers are required to write a snippet or script for this. However, this doesn't solve the problem as the script written for one problem investigation may reasonably not be applicable when the problem space alters. The lack of formal assessment environment adoption is another justified reason for this. Well, adoption of all existing reported assessment environments cannot be always encouraged as they may vary too with the demands of simulation and the type of data being used. So, a more benchmark simulation environment is needed for an IoT. This eventually gives rise to multiple arenas of research questions, e.g.

- What strategy can be adopted for modeling heterogeneous IoT devices to facilitate concurrent operation?
- What mechanism can be adopted towards achieving granularity in investigation towards the inclusion of a massive number of IoT devices over dynamic topology?
- How to develop a cost-effective simulator which has an inclusion of the majority of privilege to investigate one single platform? Unless all these local issues are not addressed, existing simulators cannot be deemed entirely reliable.

2) *Second level of investigation:* The second level of research gap is the global level extracted from the first level as a local form. This research gap is the direct consequence of the first level gap analysis. Therefore, the following are the finalized version of the research gap.

a) From all the existing features available in simulators, one potential problem is that one simulator cannot be used to carry out an extensive investigation of the issues that are not supported by it. With concurrency towards accessing infrastructure, there are inevitable complexities associated with identifying the uncertain problems that stay low and hidden while contributing to declined or unpredictable outcomes. Hence, the primary research gap is existing simulators doesn't facilitate granular investigation to identify the attributes affecting the operation of IoT device/services.

b) A closer look at the local level of open-end issues in existing simulators are i) scalability, ii) device heterogeneity, iii) concurrency towards accessibility, etc. If all these problems are looked at deeply, it can be seen that the root cause of all of the issues is the lack of adoption of resource parametric modeling in the existing simulation tool. Current simulators can perform initialization of resource parameters while the developers must script the constraints associated with its usage. Such scripts are consistently required to be upgraded with the change of services or network operations. Hence, the secondary research gap is that existing simulators are not designed to practically consider the resource modelling of IoT devices and other devices in the process of simulation. The consequences of such modeling will lead to unrealistic data transmission simulation outcomes.

c) It is closely observed that security is a much more standard set of the problem being addressed by existing simulators. However, they do it by an available set of libraries and developers' written security scripts. Almost all these security approaches are based on cryptographic based. The prime reason is its dependency on using cloud or fog as an environment on top of IoT applications, which has higher supportability towards the conventional encryption-based operation. The beneficial point in this factor is that they are 100% stopping a specific set of attackers that is coded. The limiting factor is that they are entirely not applicable if the attacker changes its plan of attack.

d) Moreover, all the cryptographic algorithms are not resistive against all attacks; they have strengths and weaknesses. Another practical rationale is that adopting cryptographic measures will also induce a higher load toward low-powered IoT devices. Hence, the ternary research gap is that existing simulators don't address the IoT nodes' sustainability factor by frequently using cryptographic measures towards security.

B. Critical Discussion for Existing Study

From the outcome of this study, it is also found that majority of the simulation tools are part of discrete event simulators (NS3, OMNeT++, SimIoT, CupCarbon, iFogSim) which offers finite set of functionalities and demands API integration. It will eventually mean that customizing them for heterogeneous research problem will be a computationally expensive process. Apart from this, existing assessment tools are mainly meant to executing standard protocol for networking and not for data analytics, which require a dependency with different set of tools. The constraints found in existing simulation tool are also associated with accessibility towards single user for one project. This will eventually pose an impediment towards distributed investigation process by different user on same project at same time. Hence, the investigation process is time consuming and platform specifics too restricted to one user at a time. Another closer observation towards existing simulation tools highlights the inferior security features embedded in it. The enhancement towards security system is very few to find, where almost all the simulation tools either uses user-deployed security patches or uses third party script to introduce security features. This process is quite challenging to be customized for projects with multiple and heterogenous target of addressing problems in IoT. Hence, there is an emergent need to develop a simulation tool that offers cost effective and proper utilization of its features.

VI. CONCLUSION

This paper has discussed the scale of effectiveness of existing simulators. There are multiple simulators in practice; the discussion has been carried out concerning commercially used and research-based usage. The novelty of this manuscript is that i) it offers an informative and compact description of frequently used simulation tools in practice for an IoT, ii) It exhibits a unique and updated research trend towards IoT simulators adoption in the last six years, which is not reported

in any existing studies, and iii) it makes some interesting discoveries of limiting factors associated with the overall features of existing simulators. The above mentioned learning outcomes of proposed review work exactly matches with the core objectives of the paper associated with studying features as well as reviewing trends associated with simulation tools. The outcome of the paper also presents open-end research problem in the form of research gap discussion, thereby meeting the core study objective of this paper.

Hence, the future work will be carried out in the direction of addressing the finalized research gap as follows:

1) The primary research gap can be addressed by developing a novel computational model of a simulator to identify and construct a set of strategies that affect the accuracy of the simulation process. Discrete mathematical modeling can be carried out to address this gap.

2) The secondary research gap can be addressed by extending the first solution toward including various novel conditional logic. The development of such reasoning is accompanied by all the resource management attributes that affect IoT devices' sustainability factors during the simulation study. It will offer different reliable outcomes due to practical resource modeling during simulation.

3) The ternary research gap can be addressed by further developing another layer of security operation without using any form of encryption model or without using any existing techniques that are found to offer load towards low-powered IoT devices. Further novelty can testify to its resiliency towards maximum forms of reported threats in an IoT.

REFERENCES

- [1] P. Zeigler, L. Zhang, Y. Iaili, "Model Engineering for Simulation," Elsevier Science, ISBN: 9780128135440, 0128135441, 2019.
- [2] D. Cvetković, G. Birajdar, Numerical Modeling and Computer Simulation, IntechOpen, ISBN: 9781838811969, 1838811966, 2020.
- [3] P. Tomar, Integration, and Implementation of the Internet of Things through Cloud Computing, Engineering Science Reference, ISBN: 9781799869832, 1799869830, 2021.
- [4] A. Khanna, D. Gupta, P. L. Mehta, V. H. C. de Albuquerque, Smart Sensors for Industrial Internet of Things Challenges, Solutions and Applications, Springer International Publishing, ISBN: 9783030526238, 3030526232, 2021.
- [5] G. Sunitha, J. Avanija, K. R. Madhavi, S. B. Bhushan, S. Goundar, Innovations in the Industrial Internet of Things (IIoT) and Smart Factory, IGI Global, ISBN: 9781799833772, 1799833771, 2021.
- [6] F. Al-Turjman, Real-Time Intelligence for Heterogeneous Networks Applications, Challenges, and Scenarios in IoT HetNets, Springer International Publishing, ISBN: 9783030756130, 3030756130, 2021.
- [7] Velayutham, Sathiyamoorthi, Challenges, and Opportunities for the Convergence of IoT, Big Data, and Cloud Computing, IGI Global, 2021.
- [8] F. Al-Turjman, Multimedia-enabled Sensors in IoT Data Delivery and Traffic Modelling, CRC Press, ISBN: 9781351166027, 1351166026, 2018.
- [9] P. Y. Taser, Emerging Trends in IoT and Integration with Data Science, Cloud Computing, and Big Data Analytics, IGI Global, ISBN: 9781799841876, 1799841871, 2021.
- [10] H. G. Perros, An Introduction to IoT Analytics, CRC Press, ISBN: 9781000337822, 1000337820, 2021.
- [11] A. Kamilaris and A. Pitsillides, "Mobile Phone Computing and the Internet of Things: A Survey," in IEEE Internet of Things Journal, vol. 3, no. 6, pp. 885-898, Dec. 2016, doi: 10.1109/JIOT.2016.2600569.

- [12] 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.
- [13] K. Tange, M. De Donno, X. Fafoutis and N. Dragoni, "A Systematic Survey of Industrial Internet of Things Security: Requirements and Fog Computing Opportunities," in *IEEE Communications Surveys & Tutorials*, vol. 22, no. 4, pp. 2489-2520, Fourth quarter 2020, doi: 10.1109/COMST.2020.3011208.
- [14] J. Zhang and D. Tao, "Empowering Things With Intelligence: A Survey of the Progress, Challenges, and Opportunities in Artificial Intelligence of Things," in *IEEE Internet of Things Journal*, vol. 8, no. 10, pp. 7789-7817, 15 May 15, 2021, doi: 10.1109/JIOT.2020.3039359.
- [15] <https://www.gambitcomm.com/site/mimic-simulator.php>.
- [16] M. Salama, Y. Elkhatib, G. Blair, "IoTNetSim: A Modelling and Simulation Platform for End-to-End IoT Services and Networking," *ACM-Proceedings of the 12th IEEE/ACM International Conference on Utility and Cloud Computing*, 2019, DOI:<https://doi.org/10.1145/3344341.3368820>.
- [17] S. Badugu, "Role of COOJA Simulator in IoT," *International Journal of Emerging Trends & Technology in Computer Science*, vol.6, Iss.2, 2017.
- [18] R. Stifani, "IBM Bluemix: The Cloud Platform for Creating and Delivering Applications," IBM Whitepaper, 2015.
- [19] <https://www.simplesoft.com/SimpleIoTSimulator.html>.
- [20] <https://www.ansys.com/en-in/technology-trends/iiot>.
- [21] <https://www.bevywise.com/iiot-simulator/>.
- [22] <https://iiotify.io/>.
- [23] C. Sonmez, A. Ozgovde, and C. Ersoy, 'Edgecloudsim: An environment for performance evaluation of edge computing systems, *Transactions on Emerging Telecommunications Technologies*, e3493, 2018.
- [24] T. Pflanzner, A. Kertesz, B. Spinnewyn and S. Latre, 'Mobiotsim: Towards a mobile iot device simulator, in 2016 IEEE 4th International Conference on Future Internet of Things and Cloud Workshops (FiCloudW), IEEE, 2016, pp. 21–27.
- [25] P. Levis and N. Lee, 'Tosim: A simulator for TinyOS networks,' UC Berkeley, September, vol. 24, 2003.
- [26] P. Gimenez, B. Molina, C. E. Palau and M. Esteve, 'Swe-simulation and testing for the iot, in 2013 IEEE International Conference on Systems, Man, and Cybernetics, IEEE, 2013, pp. 356–361.
- [27] <https://www.atomiton.com/#/home>.
- [28] <https://www.scalable-networks.com/products/qualnet-network-simulation-software/#>.
- [29] <https://www.nsnam.org/>.
- [30] <https://omnetpp.org/>.
- [31] P. Akilandeswari, B. Vennila, and H. Srimathi, "Concurrent processing of cloudlets in CloudSim-SimIoT environment," *AIP Conference Proceedings*, 2019, DOI: <https://doi.org/10.1063/1.5112277>.
- [32] K. Mehdi, M.Lounis, A. Bounceur, T. Kechadi, "CupCarbon: A Multi-Agent and Discrete Event Wireless Sensor Network Design and Simulation Tool".7th International ICST Conference on Simulation Tools and Techniques, Lisbon, Portugal,2014, DOI: 10.4108/icst.simutools.2014.254811.
- [33] X. Zeng, S. K. Garg, P. Strazdins, P. P. Jayaraman, D. Georgakopoulos and R. Ranjan, 'Iotsim: A simulator for analyzing iot applications,' *Journal of Systems Architecture*, vol. 72, pp. 93–107, 2017.
- [34] H. Gupta, A. Vahid Dastjerdi, S. K. Ghosh and R. Buyya, 'Iofogsim: A toolkit for modeling and simulation of resource management techniques in the Internet of things, edge and fog computing environments, Software: Practice and Experience, pp. 1275–1296, 2017.
- [35] S. N. Han, G. M. Lee, N. Crespi, K. Heo, N. Van Luong, M. Brut, and P. Gatellier, 'Dpwsim: A simulation toolkit for iot applications using devices profile for web services, in 2014 IEEE World Forum on Internet of Things (WF-IoT), IEEE, 2014, pp. 544–547.
- [36] M. Chernyshev, Z. Baig, O. Bello, and S. Zeadally, 'Internet of things (iot): Research, simulators, and testbeds, *IEEE Internet of Things Journal*, vol. 5, no. 3, pp. 1637–1647, 2018.
- [37] L. Sanchez, L. Munoz, J. A. Galache, P. Sotres, J. R. Santana, V. Gutierrez, R. Ramdhany, A. Gluhak, S. Krco, E. Theodoridis et al., 'Smartsantander: Iot experimentation over a smart city testbed', *Computer Networks*, vol. 61, pp. 217–238, 2014.
- [38] A. Gyrard and M. Serrano, 'Fiesta-iiot: Federated interoperable semantic internet of things (iiot) testbeds and applications,' in *ICT*, 2015.
- [39] C. Adjih, E. Baccelli, E. Fleury, G. Harter, N. Mitton, T. Noel, R. Pissard-Gibollet, F. Saint-Marcel, G. Schreiner, J. Vandaele, et al., 'Fit iot-lab: A large scale open experimental IoT testbed—a valuable tool for iot deployment in smart factories,' *IEEE ComSoc Multimedia Technical Committee E-Letter*, 2015.
- [40] J. Zhou, Z. Ji, M. Varshney, Z. Xu, Y. Yang, M. Marina, and R. Bagrodia, 'Whynet: A hybrid testbed for largescale, heterogeneous and adaptive wireless networks,' in *Proceedings of the 1st international workshop on Wireless network testbeds, experimental evaluation & characterization*, ACM, 2006, pp. 111–112.
- [41] A. Ahmad, F. Bouquet, E. Fournieret, F. Le Gall and B. Legeard, 'Model-based testing as a service for iot platforms,' in *International Symposium on Leveraging Applications of Formal Methods*, Springer, 2016, pp. 727–742.
- [42] S. Chowdhury, A. Benslimane and C. Giri, "Noncooperative Gaming for Energy-Efficient Congestion Control in 6LoWPAN," in *IEEE Internet of Things Journal*, vol. 7, no. 6, pp. 4777-4788, June 2020. doi: 10.1109/JIOT.2020.2969272.
- [43] S. Taghizadeh, H. Elbiaze, and H. Bobarshad, "EM-RPL: Enhanced RPL for Multigateway Internet-of-Things Environments," in *IEEE Internet of Things Journal*, vol. 8, no. 10, pp. 8474-8487, 15 May 15, 2021. doi: 10.1109/JIOT.2020.3047079.
- [44] M. Mahyoub, A. S. Hasan Mahmoud, M. Abu-Amara, and T. R. Sheltami, "An Efficient RPL-Based Mechanism for Node-to-Node Communications in IoT," in *IEEE Internet Things Journal*, vol. 8, no. 9, pp. 7152-7169, 1 May 1, 2021. doi: 10.1109/JIOT.2020.3038696.
- [45] I. Tomić and J. A. McCann, "A Survey of Potential Security Issues in Existing Wireless Sensor Network Protocols," in *IEEE Internet of Things Journal*, vol. 4, no. 6, pp. 1910-1923, Dec. 2017. doi: 10.1109/JIOT.2017.2749883.
- [46] Y. Kim and J. Paek, "NG-RPL for Efficient P2P Routing in Low-Power Multihop Wireless Networks," in *IEEE Access*, vol. 8, pp. 182591-182599, 2020. doi: 10.1109/ACCESS.2020.3028771.
- [47] M. Amoretti, O. Alphand, G. Ferrari, F. Rousseau, and A. Duda, "DINAS: A Lightweight and Efficient Distributed Naming Service for All-IP Wireless Sensor Networks," in *IEEE Internet of Things Journal*, vol. 4, no. 3, pp. 670-684, June 2017. doi: 10.1109/JIOT.2016.2640317.
- [48] A. W. Abbas and S. N. K. Marwat, "Scalable Emulated Framework for IoT Devices in Smart Logistics Based Cyber-Physical Systems: Bonded Coverage and Connectivity Analysis," in *IEEE Access*, vol. 8, pp. 138350-138372, 2020. doi: 10.1109/ACCESS.2020.3012458.
- [49] Kamaldeep, M. Malik, M. Dutta, and J. Granjal, "IoT-Sentry: A Cross-Layer-Based Intrusion Detection System in Standardized Internet of Things," in *IEEE Sensors Journal*, vol. 21, no. 24, pp. 28066-28076, 15 Dec.15, 2021. doi: 10.1109/JSEN.2021.3124886.
- [50] Y. Tahir, S. Yang, and J. McCann, "BRPL: Backpressure RPL for High-Throughput and Mobile IoTs," in *IEEE Transactions on Mobile Computing*, vol. 17, no. 1, pp. 29-43, 1 January 2018. doi: 10.1109/TMC.2017.2705680.
- [51] R. Monica, L. Davoli and G. Ferrari, "A Wave-Based Request-Response Protocol for Latency Minimization in WSNs," in *IEEE Internet of Things Journal*, vol. 6, no. 5, pp. 7971-7979, Oct. 2019. doi: 10.1109/JIOT.2019.2914578.
- [52] P. Sanmartin, D. Jabba, R. Sierra, and E. Martinez, "Objective Function BF-ETX for RPL Routing Protocol," in *IEEE Latin America Transactions*, vol. 16, no. 8, pp. 2275-2281, Aug. 2018 doi: 10.1109/TLA.2018.8528246.
- [53] M. Ma, W. Lin, J. Zhang, P. Wang, Y. Zhou, and X. Liang, "Toward Energy-Awareness Smart Building: Discover the Fingerprint of Your Electrical Appliances," in *IEEE Transactions on Industrial Informatics*, vol. 14, no. 4, pp. 1458-1468, April 2018. doi: 10.1109/TII.2017.2776300.

- [54] R. Bocu and C. Costache, "A homomorphic encryption-based system for securely managing personal health metrics data," in *IBM Journal of Research and Development*, vol. 62, no. 1, pp. 1:1-1:10, 1 Jan.-Feb. 2018. doi: 10.1147/JRD.2017.2755524.
- [55] Y. Shafiq, J. Henriks, C. P. Ambulo, T. H. Ware, and S. V. Georgakopoulos, "A Passive RFID Temperature Sensing Antenna With Liquid Crystal Elastomer Switching," in *IEEE Access*, vol. 8, pp. 24443-24456, 2020. doi: 10.1109/ACCESS.2020.2969969.
- [56] M. Shih, C. Huang, T. Chen, C. Wang, D. Tarn and C. P. Hung, "Electrical, Thermal, and Mechanical Characterization of eWLB, Fully Molded Fan-Out Package, and Fan-Out Chip Last Package," in *IEEE Transactions on Components, Packaging and Manufacturing Technology*, vol. 9, no. 9, pp. 1765-1775, Sept. 2019. doi: 10.1109/TCPMT.2019.2935477.
- [57] L. Fan *et al.*, "Stretchable Carbon Nanotube Thin-Film Transistor Arrays Realized by a Universal Transferable-Band-Aid Method," in *IEEE Transactions on Electron Devices*, vol. 68, no. 11, pp. 5879-5885, Nov. 2021. doi: 10.1109/TED.2021.3114140.
- [58] I. -D. Filip, F. Pop, C. Serbanescu and C. Choi, "Microservices Scheduling Model Over Heterogeneous Cloud-Edge Environments As Support for IoT Applications," in *IEEE Internet of Things Journal*, vol. 5, no. 4, pp. 2672-2681, Aug. 2018. doi: 10.1109/JIOT.2018.2792940.
- [59] S. Pang, W. Li, H. He, Z. Shan and X. Wang, "An EDA-GA Hybrid Algorithm for Multi-Objective Task Scheduling in Cloud Computing," in *IEEE Access*, vol. 7, pp. 146379-146389, 2019. doi: 10.1109/ACCESS.2019.2946216.
- [60] C. Esposito, A. Castiglione, F. Palmieri, and A. D. Santis, "Integrity for an Event Notification Within the Industrial Internet of Things by Using Group Signatures," in *IEEE Transactions on Industrial Informatics*, vol. 14, no. 8, pp. 3669-3678, Aug. 2018. doi: 10.1109/TII.2018.2791956.
- [61] A. A. Al-Roubaiey, T. R. Sheltami, A. S. H. Mahmoud and K. Salah, "Reliable Middleware for Wireless sensor-actuator Networks," in *IEEE Access*, vol. 7, pp. 14099-14111, 2019. doi: 10.1109/ACCESS.2019.2893623.
- [62] C. Esposito, M. Ficco, A. Castiglione, F. Palmieri, and A. De Santis, "Distributed Group Key Management for Event Notification Confidentiality Among Sensors," in *IEEE Transactions on Dependable and Secure Computing*, vol. 17, no. 3, pp. 566-580, 1 May-June 2020. doi: 10.1109/TDSC.2018.2799227.
- [63] K. Husain and A. Awang, "Forwarding Angles and the Trade-Off Between Reliability, Latency and Unicast Efficiency in Content-Based Beaconless Forwarding," in *IEEE Access*, vol. 8, pp. 225522-225538, 2020. doi: 10.1109/ACCESS.2020.3044967.
- [64] Z. Liu, C. Guo and B. Wang, "A Physically Secure, Lightweight Three-Factor and Anonymous User Authentication Protocol for IoT," in *IEEE Access*, vol. 8, pp. 195914-195928, 2020. doi: 10.1109/ACCESS.2020.3034219.
- [65] G. Kaur, P. Chanak and M. Bhattacharya, "Energy-Efficient Intelligent Routing Scheme for IoT-Enabled WSNs," in *IEEE Internet of Things Journal*, vol. 8, no. 14, pp. 11440-11449, 15 July 15, 2021. doi: 10.1109/JIOT.2021.3051768.
- [66] S. Banerjee, V. Odelu, A. K. Das, S. Chattopadhyay, J. J. P. C. Rodrigues and Y. Park, "Physically Secure Lightweight Anonymous User Authentication Protocol for Internet of Things Using Physically Unclonable Functions," in *IEEE Access*, vol. 7, pp. 85627-85644, 2019. doi: 10.1109/ACCESS.2019.2926578.
- [67] S. Atiewi *et al.*, "Scalable and Secure Big Data IoT System Based on Multifactor Authentication and Lightweight Cryptography," in *IEEE Access*, vol. 8, pp. 113498-113511, 2020. doi: 10.1109/ACCESS.2020.3002815.
- [68] B. Bordel, R. Alcarria, D. M. De Andrés and I. You, "Securing Internet-of-Things Systems Through Implicit and Explicit Reputation Models," in *IEEE Access*, vol. 6, pp. 47472-47488, 2018. doi: 10.1109/ACCESS.2018.2866185.
- [69] S. Dawaliby, A. Bradai, and Y. Pousset, "Distributed Network Slicing in Large Scale IoT Based on Coalitional Multi-Game Theory," in *IEEE Transactions on Network and Service Management*, vol. 16, no. 4, pp. 1567-1580, Dec. 2019. doi: 10.1109/TNSM.2019.2945254.
- [70] E. Ezenogho, K. Djuani, and A. M. Kurien, "Integrating Artificial Intelligence Internet of Things and 5G for Next-Generation Smart grid: A Survey of Trends Challenges and Prospect," in *IEEE Access*, vol. 10, pp. 4794-4831, 2022. doi: 10.1109/ACCESS.2022.3140595.
- [71] B. D. Deebak, F. Al-Turjman, M. Aloqaily, and O. Alfandi, "An Authentic-Based Privacy Preservation Protocol for Smart e-Healthcare Systems in IoT," in *IEEE Access*, vol. 7, pp. 135632-135649, 2019. doi: 10.1109/ACCESS.2019.2941575.
- [72] G. Choudhary, P. V. Astillo, I. You, K. Yim, I. -R. Chen and J. -H. Cho, "Lightweight Misbehavior Detection Management of Embedded IoT Devices in Medical Cyber-Physical Systems," in *IEEE Transactions on Network and Service Management*, vol. 17, no. 4, pp. 2496-2510, Dec. 2020. doi: 10.1109/TNSM.2020.3007535.
- [73] S. Messaoud, A. Bradai, and E. Moulay, "Online GMM Clustering and Mini-Batch Gradient Descent Based Optimization for Industrial IoT 4.0," in *IEEE Transactions on Industrial Informatics*, vol. 16, no. 2, pp. 1427-1435, Feb. 2020. doi: 10.1109/TII.2019.2945012.
- [74] J. Yang, A. S. Akyurek, S. Tilak and T. S. Rosing, "Design of Transmission Manager in Heterogeneous WSNs," in *IEEE Transactions on Emerging Topics in Computing*, vol. 6, no. 3, pp. 395-408, 1 July-Sept. 2018. doi: 10.1109/ETC.2017.2653064.
- [75] C. Pu, "Sybil Attack in RPL-Based Internet of Things: Analysis and Defenses," in *IEEE Internet of Things Journal*, vol. 7, no. 6, pp. 4937-4949, June 2020. doi: 10.1109/JIOT.2020.2971463.
- [76] E. A. Khalil, S. Ozdemir and B. A. Attea, "A New Task Allocation Protocol for Extending Stability and Operational Periods in the Internet of Things," in *IEEE Internet of Things Journal*, vol. 6, no. 4, pp. 7225-7231, Aug. 2019. doi: 10.1109/JIOT.2019.2915558.
- [77] C. Pu and L. Carpenter, "Psched: A Priority-Based Service Scheduling Scheme for the Internet of Drones," in *IEEE Systems Journal*, vol. 15, no. 3, pp. 4230-4239, Sept. 2021. doi: 10.1109/JSYST.2020.2998010.
- [78] A. Ali and M. M. Yousaf, "Novel Three-Tier Intrusion Detection and Prevention System in Software Defined Network," in *IEEE Access*, vol. 8, pp. 109662-109676, 2020. doi: 10.1109/ACCESS.2020.3002333.
- [79] T. Qayyum, Z. Trabelsi, A. W. Malik and K. Hayawi, "Multi-Level Resource Sharing Framework Using Collaborative Fog Environment for Smart Cities," in *IEEE Access*, vol. 9, pp. 21859-21869, 2021. doi: 10.1109/ACCESS.2021.3054420.
- [80] B. Pan, F. Yan, X. Xue, E. Magelhaes and N. Calabretta, "Performance assessment of a fast optical add-drop multiplexer-based metro access network with edge computing," in *Journal of Optical Communications and Networking*, vol. 11, no. 12, pp. 636-646, December 2019. doi: 10.1364/JOCN.11.000636.
- [81] A. Arooj, M. S. Farooq, T. Umer, G. Rasool, and B. Wang, "Cyber-Physical and Social Networks in IoV (CPSN-IoV): A Multimodal Architecture in Edge-Based Networks for Optimal Route Selection Using 5G Technologies," in *IEEE Access*, vol. 8, pp. 33609-33630, 2020. doi: 10.1109/ACCESS.2020.2973461.
- [82] A. Alharthi, Q. Ni and R. Jiang, "A Privacy-Preservation Framework Based on Biometrics Blockchain (BBC) to Prevent Attacks in VANET," in *IEEE Access*, vol. 9, pp. 87299-87309, 2021. doi: 10.1109/ACCESS.2021.3086225.
- [83] M. S. Akbar, H. Yu, and S. Cang, "Performance Optimization of the IEEE 802.15.4-Based Link Quality Protocols for WBANs/IoTs in a Hospital Environment Using Fuzzy Logic," in *IEEE Sensors Journal*, vol. 19, no. 14, pp. 5865-5877, 15 July 15, 2019. doi: 10.1109/JSEN.2019.2900009.
- [84] H. Huang, F. Liu, Z. Yang, and Z. Hao, "Automated Test Case Generation Based on Differential Evolution With Relationship Matrix for iFogSim Toolkit," in *IEEE Transactions on Industrial Informatics*, vol. 14, no. 11, pp. 5005-5016, Nov. 2018. doi: 10.1109/TII.2018.2856881.
- [85] K. S. Awaisi *et al.*, "Towards a Fog Enabled Efficient Car Parking Architecture," in *IEEE Access*, vol. 7, pp. 159100-159111, 2019. doi: 10.1109/ACCESS.2019.2950950.
- [86] F. H. Rahman, S. H. S. Newaz, T. W. Au, W. S. Suhaili and G. M. Lee, "Off-Street Vehicular Fog for Catering Applications in 5G/B5G: A Trust-Based Task Mapping Solution and Open Research Issues," in

- IEEE Access*, vol. 8, pp. 117218-117235, 2020. doi: 10.1109/ACCESS.2020.3004738.
- [87] H. Nashaat, E. Ahmed and R. Rizk, "IoT Application Placement Algorithm Based on Multi-Dimensional QoE Prioritization Model in Fog Computing Environment," in *IEEE Access*, vol. 8, pp. 111253-111264, 2020. doi: 10.1109/ACCESS.2020.3003249.
- [88] B. K. Dar, M. A. Shah, S. U. Islam, C. Maple, S. Mussadiq and S. Khan, "Delay-Aware Accident Detection and Response System Using Fog Computing," in *IEEE Access*, vol. 7, pp. 70975-70985, 2019. doi: 10.1109/ACCESS.2019.2910862.
- [89] H. Rafique, M. A. Shah, S. U. Islam, T. Maqsood, S. Khan and C. Maple, "A Novel Bio-Inspired Hybrid Algorithm (NBIHA) for Efficient Resource Management in Fog Computing," in *IEEE Access*, vol. 7, pp. 115760-115773, 2019. doi: 10.1109/ACCESS.2019.2924958.
- [90] M. Ammad *et al.*, "A Novel Fog-Based Multi-Level Energy-Efficient Framework for IoT-Enabled Smart Environments," in *IEEE Access*, vol. 8, pp. 150010-150026, 2020. doi: 10.1109/ACCESS.2020.3010157.
- [91] I. Lera, C. Guerrero and C. Juiz, "YAFS: A Simulator for IoT Scenarios in Fog Computing," in *IEEE Access*, vol. 7, pp. 91745-91758, 2019. doi: 10.1109/ACCESS.2019.2927895.
- [92] J. U. Arshed and M. Ahmed, "RACE: Resource Aware Cost-Efficient Scheduler for Cloud Fog Environment," in *IEEE Access*, vol. 9, pp. 65688-65701, 2021. doi: 10.1109/ACCESS.2021.3068817.
- [93] R. Yadav *et al.*, "Smart Healthcare: RL-Based Task Offloading Scheme for Edge-Enable Sensor Networks," in *IEEE Sensors Journal*, vol. 21, no. 22, pp. 24910-24918, 15 Nov.15, 2021. doi: 10.1109/JSEN.2021.3096245.
- [94] B. Ali, M. Adeel Pasha, S. u. Islam, H. Song and R. Buyya, "A Volunteer-Supported Fog Computing Environment for Delay-Sensitive IoT Applications," in *IEEE Internet of Things Journal*, vol. 8, no. 5, pp. 3822-3830, 1 March 1, 2021. doi: 10.1109/JIOT.2020.3024823.
- [95] J. Fang and A. Ma, "IoT Application Modules Placement and Dynamic Task Processing in Edge-Cloud Computing," in *IEEE Internet of Things Journal*, vol. 8, no. 16, pp. 12771-12781, 15 August 15, 2021. doi: 10.1109/JIOT.2020.3007751.
- [96] A. Asghar, A. Abbas, H. A. Khattak, and S. U. Khan, "Fog Based Architecture and Load Balancing Methodology for Health Monitoring Systems," in *IEEE Access*, vol. 9, pp. 96189-96200, 2021. doi: 10.1109/ACCESS.2021.3094033.
- [97] S. N. Han *et al.*, "DPWSim: A Devices Profile for Web Services (DPWS) Simulator," in *IEEE Internet of Things Journal*, vol. 2, no. 3, pp. 221-229, June 2015. doi: 10.1109/JIOT.2014.2388131.
- [98] S. N. Han, G. M. Lee and N. Crespi, "Semantic Context-Aware Service Composition for Building Automation System," in *IEEE Transactions on Industrial Informatics*, vol. 10, no. 1, pp. 752-761, Feb. 2014. doi: 10.1109/TII.2013.2252356.
- [99] M. S. Al-Rakhami and M. Al-Mashari, "ProChain: Provenance-Aware Traceability Framework for IoT-Based Supply Chain Systems," in *IEEE Access*, vol. 10, pp. 3631-3642, 2022, doi: 10.1109/ACCESS.2021.3135371.
- [100] A. S. M. S. Hosen, P. K. Sharma and G. H. Cho, "MSRM-IoT: A Reliable Resource Management for Cloud, Fog, and Mist-Assisted IoT Networks," in *IEEE Internet of Things Journal*, vol. 9, no. 4, pp. 2527-2537, 15 Feb.15, 2022, doi: 10.1109/JIOT.2021.3090779.
- [101] T. Kamalakis, Z. Ghassemlooy, S. Zvanovec and L. Nero Alves, "Analysis and simulation of a hybrid visible-light/infrared optical wireless network for IoT applications," in *Journal of Optical Communications and Networking*, vol. 14, no. 3, pp. 69-78, March 2022, doi: 10.1364/JOCN.442787.
- [102] X. Chen, J. Zhang, B. Lin, Z. Chen, K. Wolter and G. Min, "Energy-Efficient Offloading for DNN-Based Smart IoT Systems in Cloud-Edge Environments," in *IEEE Transactions on Parallel and Distributed Systems*, vol. 33, no. 3, pp. 683-697, 1 March 2022, doi: 10.1109/TPDS.2021.3100298.
- [103] J. A. Barriga, P. J. Clemente, J. Hernández and M. A. Pérez-Toledano, "SimulateIoT-FIWARE: Domain Specific Language to Design, Code Generation and Execute IoT Simulation Environments on FIWARE," in *IEEE Access*, vol. 10, pp. 7800-7822, 2022, doi: 10.1109/ACCESS.2022.3142894.

Analysis and Prediction of COVID-19 by using Recurrent LSTM Neural Network Model in Machine Learning

N.P.Dharani¹

Research Scholar, Department of Electronics and
Communication Engineering, Koneru Lakshmaiah
Education Foundation, Guntur, India
Sree Vidyaniethan Engineering College
Tirupati, Andhra Pradesh, India

PolaiahBojja²

Professor, Institute of Aeronautical Engineering
Hyderabad, TS, India
Koneru Lakshmaiah Education Foundation
Guntur, India

Abstract—As we all know that corona virus is announced as pandemic in the world by WHO. It is spreaded all over the world with few days of time. To control this spreading, every citizen maintains social distance and self preventive measures are the best strategies. As of now many researchers and scientists are continuing their research in finding out the exact vaccine. The machine learning model finds that the corona virus disease behaves in exponential manner. To abolish the consequence of this pandemic, an efficient step should be taken to analyze this disease. In this paper, a recurrent neural network model is chosen to predict the number of active cases in a particular state. To do this prediction of active cases, we need database. The database of COVID-19 is downloaded from KAGGLE website and is analyzed by applying recurrent LSTM neural network with univariant features to predict for the number of active cases of patients suffering from corona virus. The downloaded database is divided into training and testing the chosen neural network model. The model is trained with the training data set and tested with testing dataset to predict the number of active cases in a particular state here we have concentrated on Andhra Pradesh state.

Keywords—COVID-19; corona virus; KAGGLE; LSTM neural network; machine learning

I. INTRODUCTION

One of the journals Nature reported that these viruses are derived from unrecognized group called as corona virus and is identified by electron microscope. The full name of COVID-2019 is the Coronavirus disease of 2019, which has created panic in the whole world today. The total population of Andhra Pradesh in the year 2020 is 91,717,240 the people affected with the disease i.e., the confirmed cases are 4,45,139, the active cases of novel corona virus are 1,01,210 as of 2nd September, 2020 which has been taken from publicly available database [1]. All most all countries announced their states to lockdown in order to stop the travel of their citizens unnecessarily. Somehow the spreading of virus is controlled due to announcement of lockdown otherwise the spread of the disease is anonymous. Even though the economy of many countries was drastically dropped, the government announced lockdown. If anyone is found to be infected, she/he will be under quarantine for 14 days and treatment is given for

recovery. Base on the condition, it may cause death and many people gone to depression level. In India, the outburst of virus is disturbed the whole functioning life. At the starting stage, the cases are increased by transmission through local i.e., from person to person and later it is continued as the same [1]. The ways to detect corona virus by using rapid test kit, a portable device also detects virus in mucus membrane using a chip and a scanner and by taking a swab sample from patient's mouth or nose.

Till now there is no correct vaccine and anti-viral treatments are available and many medical organizations are trying hard to find out vaccine for COVID-19[2]. It is in our hands to save our lives from corona virus by providing personal protective equipments, masks and sanitization and maintaining social distance [3]. If we consider the present situation of COVID-19, the qualitative information is more prominent when compare with quantitative information. A best suited mathematical model is not able to predict the whole disease but, it may study the model to derive the nature of the disease. So, an appropriate machine learning/deep learning models are best suited to predict and study the nature and behavior of the whole disease shortly [4][5].

Artificial neural networks are very similar to our biological learning system that is interconnected with many several neurons in brain. ANN systems are provoked to confine this type of parallel computation based on distribution representations. To generate single valued output from real valued inputs, ANNs should be set up with a densely interconnected simple set of units. Here, interconnection is simply expressed as the means of processing the elements in neural network which are interconnected to one another [6]. So, the provision of all the elements and structure of connections are important in artificial neural network. Normally, we have three layers in ANN system. One is input layer where the inputs are feeding to the network and output layer, which generates outputs based on the inputs that we have provided to the system. The last and important layer is hidden layer where it acts as an interface between the input and the output layers. If we keep on increasing the hidden layers, the power required to process and computational speed can be increased and the entire system become complex.

Another class of ANN is recurrent neural network (RNN) [7]. The connections are formed between nodes and by a directed graph all along a series. This forms a dynamic behavior for a time series. These networks have feedback and form a closed loop. RNN also uses memory to process the series of inputs that we provided to the network. RNN have single layered recurrent network and multi layered recurrent network.

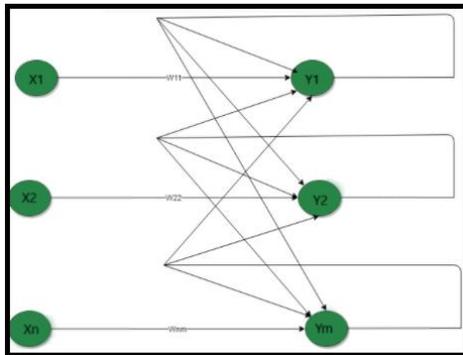


Fig. 1. Single Layered Recurrent Neural Network Model.

Fig. 1 is a network that represents a single layer network which provides a feedback connection in which each element of the node is given feedback to its own element or other element or can be to both.

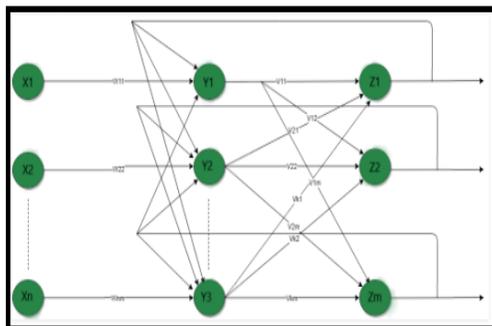


Fig. 2. Multi Layered Recurrent Neural Network Model.

Here, in Fig. 2, the multilayered network is shown; the output of the element is directed to the other element of the same layer and to the previous layer which forms a multi RNN. Both the elements perform the same operation and the output depends on the previous calculations so here no need to have inputs at each step. The computations of the series are captured in the hidden layer.

II. RELATED WORK

Lin Jia et al. analyzed three different types of mathematical models namely Bertalanffy, Gompertz and Logistic models. They applied these three models for different regions and found different parameters. With these parameters, they found Logistic model gives the outer performance among all the three models [7]. Narinder Singh Punn developed mathematical models like SVR, DNN, RNN and PR to find the RMSE and concluded that PR model gives the less RMSE value when compare with other models [8]. Sarbjit Singh et al. developed a hybrid model which involves

the decomposition of dataset into series of components by applying discrete wavelet function and then these components are applied to a suitable model named as ARIMA model for prediction of death cases for next one month across five countries [9].

LinhaoZhong et al. proposed a mathematical model for early prediction of number of infected cases by using SIR model with minimum parameters like recovered rate and infectious rate. Since the number of cases are exponentially increasing manner, quarantine measures need to be followed strictly and must pay attention towards the medical service [10]. UtkucanSahin et al. presented a model to forecast the number of confirmed cases in UK, USA and Italy. The authors studied a nonlinear grey Bernoulli model, grey model and fractional nonlinear grey Bernoulli model to predict confirmed cases. In their study fractional nonlinear grey Bernoulli model offers the best performance of providing lowest MAPE, RMSE and R^2 values [11]. Lixiang Li et al. proposed suitable model to compare the official data and model predicted data and found the error is very small [12]. NaliniChintalapudi et al. presented a model using R statistics to forecast the registered and recovered cases [13]. Debanjan Parbat et. al. utilized SVR model for prediction of total number of recovered, death, confirmed cases and found the accuracy of the model along with MSE and RMSE [14]. SalihDjilaliet.al. presented a mathematical model to predict the spread of the disease transmission [15]. Patricia Melinet. al. proposed a neural network multi ensemble model with fuzzy response for the corona virus time series data to get valid and accurate predicted values [16]. Zlatan Car, Sandi BaressiSegota et al proposed a neural network model of multilayer perceptron to find R^2 values for recovered, confirmed and death cases [17].

III. DATASETS AND METHODOLOGY

A. Dataset Description

In this analysis, the COVID-19 data was downloaded from KAGGLE site. There are different source to get the data for analysis which includes: (1) john Hopkins (<https://corona.virus.jhu.edu/>); (2) KAGGLE (<https://www.kaggle.com/sudalairajkumar/covid19-in-india>); (3) CDC (<https://www.cdc.gov/library/researchguides/2019novelcoronavirus/researcharticles.html>); (4) data hub (<https://datahub.io/core/covid-19>); (5) Tableau (<https://www.tableau.com/covid-19-coronavirus-data-resources>) and soon. With these websites any researcher can download the datasets which is his/her interest and can do analysis. The data we have taken from 30 January, 2020 to 2 September 2020 which consists of confirmed, death, cured/migrated cases of all over India. But in this research article, we concentrated on only the state Andhra Pradesh. The dataset of daily reported cases are summarized in a table in the form of XLSX or CSV format with the parameters like confirmed, deaths, cured/migrated cases. These datasets are taken in this paper for analysis and prediction of active cases especially in Andhra Pradesh.

B. Methodology

The corona virus in India spreads due to local transmission from one person to other person easily at the earliest stage.

The expert person has to diagnose at the earliest stage and can control the spread of the disease. With the objective of forecasting the possibility of transmission among citizens, we developed a recurrent neural network model. This system model utilizes long short term memory (LSTM) cell [18]. To develop this, machine learning and deep learning library packages like pandas, numpy, matplotlib, seaborn, sklearn and math are imported into jupyter notebook to analyze confirmed, death cases and also to predict active cases of Andhra Pradesh (AP).

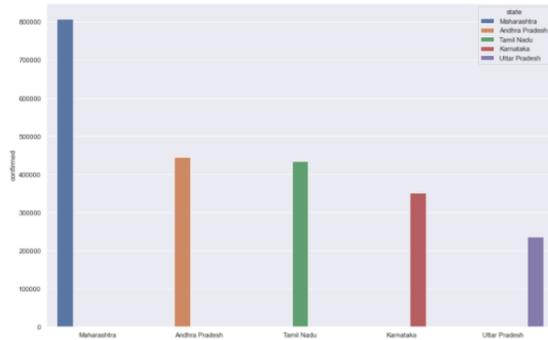


Fig. 3. Bar Plot of Top 5 States of Confirmed Cases.

The plot shown in Fig. 3 provides the information of confirmed cases where Maharashtra is in top position, Andhra Pradesh stood second position and fifth place is Uttar Pradesh. Similarly, if we consider top death cases, the below Fig. 4 shows the bar pot of states with top death cases.

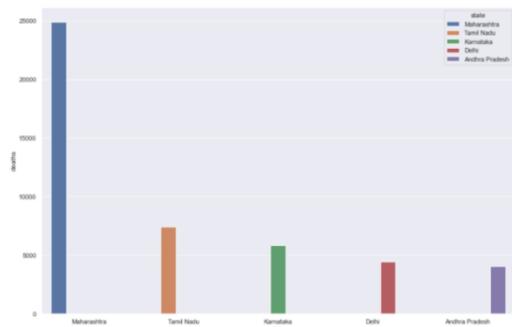


Fig. 4. Bar Plot of States with Top Death Cases.

Now if we consider the total number of confirmed cases, death cases and active cases in Andhra Pradesh. Plot the line graph with the data available in CSV file, we get the graphs that was shown in Fig. 5, 6 and 7, respectively:

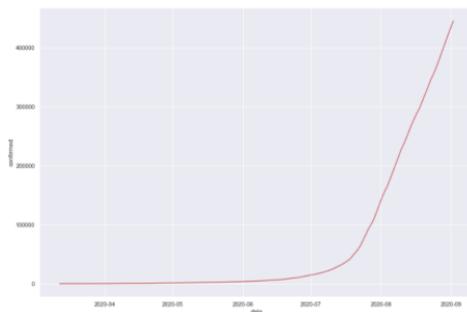


Fig. 5. Plot of Confirmed Cases in AP.

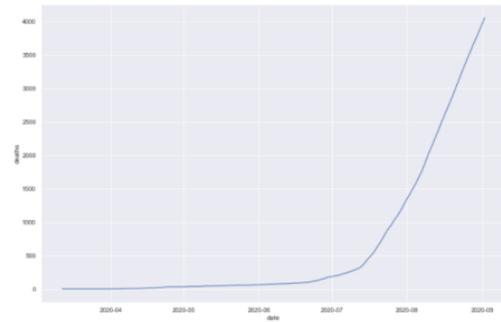


Fig. 6. Plot of Deaths Cases in AP.

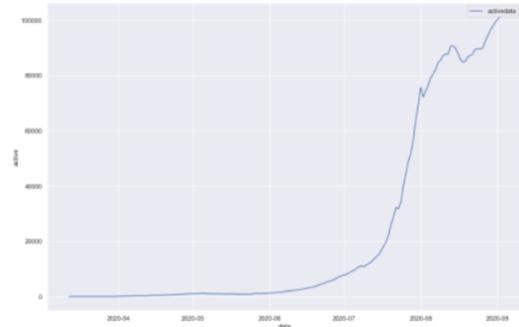


Fig. 7. Plot of Active Cases in AP.

IV. IMPLEMENTATION OF ALGORITHM

A conventional neural network is having just a bunch of parallel layers which consists of nodes called as neurons. These neurons are interconnected to each other and forms layers from which data transmits from one layer to the next layer. This first layer is named as input layer, last layer is called as output layer and between layers are named as hidden layers [19]. A set of neural networks known as recurrent networks which deals with the time series data. These networks have memory to process the previous data that is transferred through the network. But the RNN experiences from short term memory. While computing gradients during back propagation, when the gradients become very small and they will not add up large amount of learning. Therefore RNN stops learning since they get very small gradients. Hence these RNN when it is seen in longer series networks won't learn much and thus have very short memory. To eliminate this short memory, LSTM should incorporate a mechanism inside the network known as gates and controls the stream of data that pass through the nodes and thus eliminates the short memory. The implementation of corona virus forecasting is based on the long short-term memory (LSTM) networks by taking one feature into account at a time [20] [21].



Fig. 8. LSTM Single Time Series Feature Model.

Fig. 8 represents an LSTM model which is a type of RNN and especially used to predict the time series patterns as well as classification problems. Since our dataset is time series, we have chosen this model for prediction. This model may take last 8 day's features and forecasts the figures for the 8th day. When the target value reaches, it stops and exists; it takes into account its predictions.

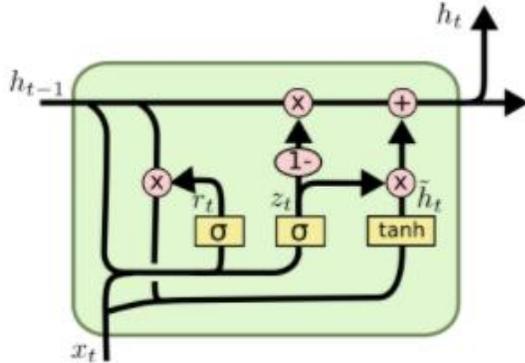


Fig. 9. LSTM Model with its Activation Function.

The LSTM model shown in Fig. 9 merges the forget gate and input gate into one gate. It also joins the cell state and hidden states. The number of nodes or neurons is chosen in trial and error method to give best results. The most common method is k-fold cross validation. The formula is expressed below to find number of nodes:

$$N_h = \frac{2}{3}(N_i + N_o) \quad (1)$$

Where N_i is the input neurons and N_o is the output neurons.

The outputs of the respective gates are given below:

$$\begin{aligned} Z_t &= \sigma(W_z \cdot [h_{t-1}, X_t]) \\ r_t &= \sigma(W_r \cdot [h_{t-1}, X_t]) \\ \tilde{h}_t &= \tanh(W \cdot [r_t * h_{t-1}, X_t]) \\ h_t &= (1 - z_t) * h_{t-1} + z_t * \tilde{h}_t \end{aligned} \quad (2)$$

1) Procedure of Neural Network1:

- a) Import required library packages.
- b) Read data which is saved in .csv format.
- c) Filter the data by choosing required state here Andhra Pradesh has been chosen.
- d) Line plot of confirmed cases of AP.
- e) Line plot of death cases.
- f) Divide the train, test data and also define number of epochs, batch size, number nodes, activation function and optimizer.
- g) Train the network model by using fit function and plot the graph of training and predicted values.
- h) Find out the performance metric like RMSE, MSE, MAE, SSE.

The loss of one layer in LSTM model is very high. To get best output, we need to add one more layer by adding nn2.add instruction.

2) Procedure of Neural Network Layer2

- a) Add the model as sequential.
- b) Add one more layer and one dense layer and repeat the compilation of nn2 to find the performance metrics.
- c) Here also all the metrics shows higher values.
- d) To reduce the metrics values, we have to convert the time series to stationary since the input we applied is dynamic time series and is not repeating.

3) Procedure to Find the Difference:

- a) Find the series of difference for the given time series as: $d_1 = L_1 - L_0, d_2 = L_2 - L_1, \dots, d_N = L_N - L_{N-1}$.
- b) From there, find $L_1 = L_0 + d_1, L_2 = L_1 + d_2, \dots, L_N = L_{N-1} + d_N$.
- c) Find cumulative series by finding $d_k = L_k - L_0, d_{k+1} = L_{k+1} - L_1, \dots, d_N = L_N - L_{N-k}$.
- d) Plot the graph of difference active cases.
- e) Again find the difference of difference of difference of active cases since, the difference we calculated in step 3 may be iterative.
- f) Compute the metrics of difference of difference of difference of active cases.
- g) Plot the graph of difference of difference of difference of active cases.

4) Procedure to Find the Scaling Transformation:

- a) First reduce the dataset into fixed interval.
- b) Use the activation function linear from 0 to 1 which eliminates negative values.
- c) Apply scaling transformation to the difference of active cases.
- d) Apply this value to third layer of LSTM model to train and to find performance metrics.
- e) Plot the actual series of active cases.

V. SIMULATION RESULTS

A. Performance Analysis

The performance parameters are analyzed to evaluate the neural network model called Mean Absolute Error, Mean Square Error, Root Mean Square Error and Sum of Squared Error and their definitions along with equations are shown below. The definitions of the performance metrics are shown along with equations [22]:

1) *Mean absolute error*: The ratio of sum of differences of all predicted values and tested values to the total length of tested values. The formula to compute MAE is given below [22] [23] [26]:

$$MAE = \frac{\sum_{N=1}^n |Actual - Predicted|}{N} \quad (3)$$

2) *Mean square error*: The mean of difference squared between the actual and predicted values. It is expressed as [22][24]:

$$MSE = \frac{\sum_{N=1}^n |Actual - Predicted|^2}{N} \quad (4)$$

3) *Root mean square error*: Root mean square error is used to compute the residual errors. These residual are measured by finding how the values are away from the line of regression. RMSE is the ratio of square root of summation of squared deviation of actual and predicted values to the number of actual values [25].

$$RMSE = \sqrt{\frac{\sum_{N=1}^n |Actual - Predicted|^2}{N}} \quad (5)$$

4) *Sum of square error*: It is defined as the sum of difference squared between all the actual and predicted values. The equation of sum of squared error is given by the formula [23]:

$$SSE = \sum_{N=1}^n |Actual - Predicted|^2 \quad (6)$$

The neural network model is trained with appropriate dataset to predict the active case in AP and computed the performance metrics.

B. Graphical Representation of Active Cases in Andhra Pradesh

Fig. 10 represents the data and predictions of active cases in AP.

Fig. 11 plots the data of predictions of active cases that was taken from second neural network model. Fig. 12 provides information of data after taking difference of active cases.



Fig. 10. Plot of Data and Predictions of Active Cases in AP.

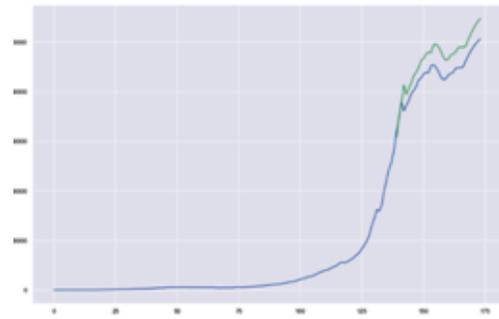


Fig. 11. Plot of Predictions of Active Cases from Second Neural Network Layer.

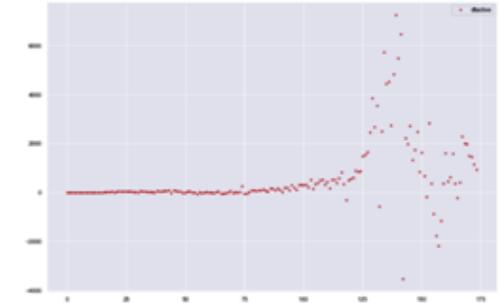


Fig. 12. Plot of Difference of Active Cases in AP State.

Fig. 13 provides information of 3 differences in active cases and analyzed that data has now become statistic instead of dynamic.

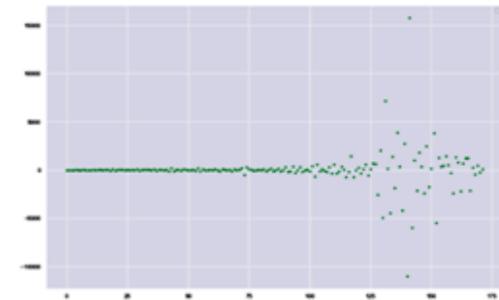


Fig. 13. Plot of 3 Differences of Active Cases in AP.

Fig. 14 represents graph of training and predicting data of active cases in Andhra Pradesh.

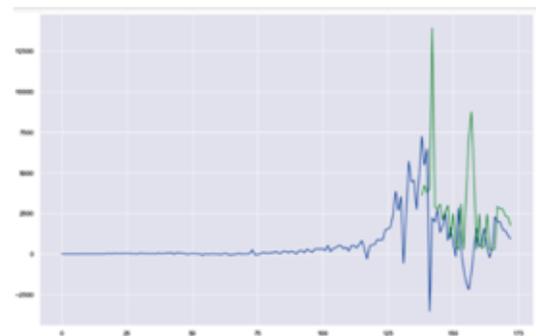


Fig. 14. Plot of Loss for Training Data and Predicting Data of Active Cases in AP.

Fig. 15 shows the actual series of the active cases in Andhra Pradesh.

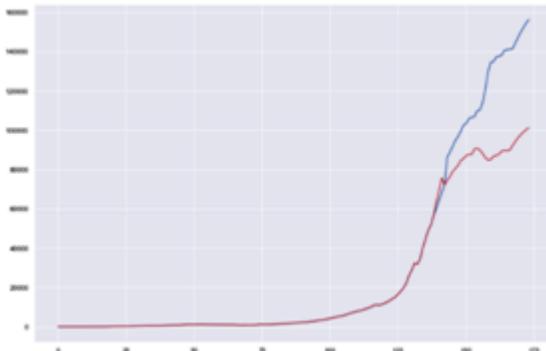


Fig. 15. Plot the Actual Series of Active Cases in AP.

Fig. 16 represents the active cases after scaling down to reduce the loss.

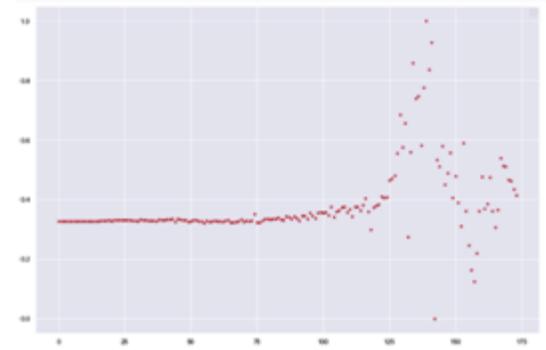


Fig. 16. Plot of Active Cases after Difference Scaling.

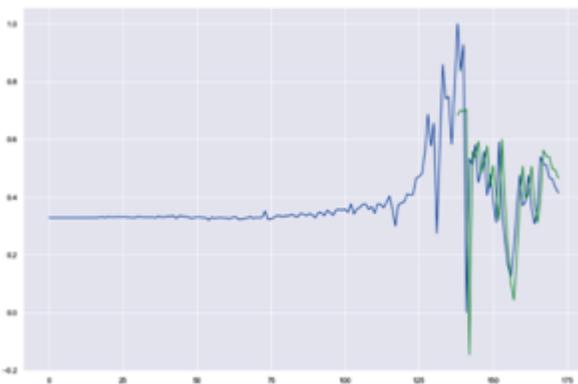


Fig. 17. Plot of Training Data and Predicting Data of Difference Scaling of Active Cases in AP.

Fig. 17 represents the information of training and predicting data of difference scaling of active case in Andhra Pradesh state.

Fig. 18 shows the actual series of cases and the predicted cases and provides the information of deviation from actual to prediction values.

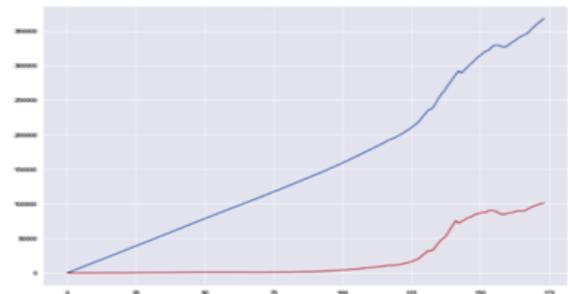


Fig. 18. Plot of Actual Series of Cumulated and Predicted Active Cases in AP.

All the performance metric values are tabulated in Table I.

TABLE I. SUMMARY OF PERFORMANCE METRIC VALUES OF THE NEURAL NETWORK MODEL

S. No	Performance metrics	Values
One layer in LSTM model		
1.	Loss	49385091.59
2.	RMSE	3151.79
3.	MAE	1187.86
4.	SSE	1728478205.76
Second layer of NN		
6.	Loss	50219150.41
7.	RMSE	3178.29
8.	MAE	1197.85
9.	SSE	1757670264.37
Difference of active cases		
11.	Loss	18699324.54
12.	RMSE	1945.02
13.	MAE	730.94
14.	SSE	654476359.02
Third layer of NN (Scaling Transformation)		
16.	Loss	0.04
17.	RMSE	0.09
18.	MAE	0.03
19.	SSE	1.49

We tabulated the actual values and predicted values of active cases in Andhra Pradesh in Table II for the next 25 days.

TABLE II. ACTUAL AND PREDICTED VALUES OF ACTIVE CASES

S.No.	Actual Values	Predicted Values	S.No.	Actual Values	Predicted Values
1.	1	1	2.	1	1453
3.	1	2907	4.	1	4359
5.	1	5812	6.	1	7265
7.	2	8719	8.	3	10172
9.	3	11625	10.	5	13080
11.	7	14535	12.	8	15989

13.	8	17441	14.	10	18896
15.	11	20350	16.	13	21805
17.	18	23262	18.	22	24719
19.	39	26189	20.	82	27686
21.	84	29140	22.	130	30640
23.	159	32122	24.	188	33604
25.	222	35091	26.	262	36584
27.	296	38072	28.	338	39567
29.	350	41032	30.	364	42499
31.	364	43952	32.	414	45455
33.	450	46944	34.	478	48425
35.	500	49900	36.	522	51375
37.	546	52853	38.	546	54305
39.	610	55823	40.	639	57305
41.	669	58788	42.	727	60300
43.	781	61807	44.	859	63339
45.	835	64767	46.	911	66297
47.	970	67810	48.	1014	69307
49.	1051	70797	50.	1027	72226
51.	1027	73678	52.	1062	75167
53.	1090	76648	54.	1092	78103
55.	1092	79555	56.	1012	80927
57.	1029	82397	58.	1004	83824
59.	999	85272	60.	1010	86736
61.	998	88177	62.	988	89619
63.	948	91032	64.	965	92502
65.	1007	93997	66.	953	95395
67.	901	96795	68.	872	98218
69.	859	99658	70.	909	101161
71.	885	102590	72.	891	104049
73.	892	105503	74.	911	106975
75.	1158	108677	76.	1105	110077
77.	1056	111480	78.	1067	112944
79.	1150	114480	80.	1220	116004
81.	1268	117505	82.	1341	119032
83.	1413	120558	84.	1546	122145
85.	1613	123666	86.	1654	125160
87.	1817	126778	88.	1951	128367
89.	2031	129900	90.	2191	131515
91.	2292	133070	92.	2301	134532
93.	2495	136181	94.	2688	137830
95.	2765	139360	96.	3052	141104
97.	3244	142751	98.	3340	144301
99.	3637	146055	100.	3948	147822
101.	4240	149571	102.	4562	151350
103.	4766	153009	104.	5284	154986
105.	5428	156585	106.	5760	158374
107.	6145	160217	108.	6648	162179
109.	7164	164154	110.	7479	165926
111.	7897	167802	112.	8071	169431

113.	8586	171406	114.	9096	173375
115.	9473	175210	116.	10043	177240
117.	10860	179520	118.	11200	181317
119.	10894	182460	120.	11383	184408
121.	11936	186421	122.	12533	188478
123.	13428	190837	124.	14274	193147
125.	15144	195481	126.	16621	198430
127.	18159	201440	128.	19814	204569
129.	22260	208499	130.	26118	213859
131.	28800	218028	132.	32336	223062
133.	31763	223935	134.	34272	227929
135.	39990	235173	136.	44431	241123
137.	48956	247159	138.	51701	251392
139.	56527	257733	140.	63771	263017
141.	69252	268331	142.	75720	273704
143.	72188	279086	144.	74404	275554
145.	76377	279574	146.	79104	283388
147.	80426	287791	148.	82166	290972
149.	84654	294572	150.	85486	298805
151.	87112	301436	152.	87773	304926
153.	87597	307352	154.	90425	308678
155.	90780	313147	156.	89907	315187
157.	88138	315508	158.	85945	314478
159.	84777	312813	160.	85130	312692
161.	86725	314730	162.	87177	318190
163.	87803	320354	164.	89389	322736
165.	89742	326188	166.	89516	328225
167.	89932	329481	168.	92208	331600
169.	94209	335669	170.	96191	339507
171.	97681	343329	172.	99129	346684
173.	100276	349996	174.	101210	352988

VI. CONCLUSION

Corona virus pandemic occurred in all over the world. By applying LSTM neural network model, we predicted the growth of active cases of Andhra Pradesh. If we observe the plot of actual series and predicted series, it is still showing exponential behavior. Every citizen need to follow preventive measures to avoid and controlling the spread of virus. This analysis shows the predicted values and performance metrics like MAE, MSE, RMSE and SSE values. Up on observation among all the mentioned metrics the third layer offers minimum error values with the LSTM model of 3 layers with 10 nodes. If we increase computing of difference between the active cases, we get the errors to minimum value.

VII. DATA AVAILABILITY

The data used to support the findings of this study are available from the corresponding author upon request.

VIII. FUNDING

This research has been no funds received for this research work.

IX. CONSENT

Informed consent was obtained from all individual participants included in the study.

X. CONFLICTS OF INTEREST

The authors declare that they have no conflicts of interest. Authors' Contributions All authors equally contributed to the study conception and design and implementation of the research, analysis and interpretation of results, and manuscript preparation.

ACKNOWLEDGMENT

The authors are thanks to Koneru Lakshmaiah Education Foundation, Guntur for providing the support and Characterization of completing this work.

REFERENCES

- [1] R. Sujath, et al., A machine learning forecasting model for COVID-19 pandemic in India, *Stochastic Environmental Research and Risk Assessment*, <https://doi.org/10.1007/s00477-020-01827-8>, May, 2020.
- [2] Organization W.H.O, et al. Q&A on coronaviruses. 2020b.
- [3] Parul Arora, et al., Prediction and analysis of COVID-19 positive cases using deep learning models: A descriptive case study of India, *Chaos, Solitons and Fractals*, 139, 2020.
- [4] Ramjeet Singh Yadav, Data analysis of COVID-2019 epidemic using machine learning methods: a case study of India, *International Journal of Information Technology*, <https://doi.org/10.1007/s41870-020-00484-y>, May, 2020.
- [5] ManotoshMandal, et al., A model based study on the dynamics of COVID-19: Prediction and control, *Chaos, Solitons and Fractals*, 136, May, 2020.
- [6] Najmul Hasan, A Methodological Approach for Predicting COVID-19 Epidemic Using EEMD-ANN Hybrid Model, *Internet of Things*, 11, May, 2020.
- [7] Jia, YuKang, et al. Long Short-Term Memory Projection Recurrent Neural Network Architectures for Piano's Continuous Note Recognition. *Journal of Robotics*, 2017.
- [8] Narinder Singh Punn, et al., COVID-19 Epidemic Analysis using Machine Learning and Deep Learning Algorithms, *Medrxiv*, 2020.
- [9] Sarbjit Singh, et al., Development of new hybrid model of discrete wavelet decomposition and autoregressive integrated moving average (ARIMA) models in application to one month forecast the casualties cases of COVID-19, *Chaos, Solitons and Fractals*, 135, 2020.
- [10] Linhao Zhong, et al., Early Prediction of the 2019 Novel Coronavirus Outbreak in the Mainland China Based on Simple Mathematical Model, *IEEE access*, vol. 8, 2020.
- [11] Utkucan Sahin and TezcanSahin, Forecasting the cumulative number of confirmed cases of COVID-19 in Italy, UK and USA using fractional nonlinear grey Bernoulli model, *Chaos, Solitons and Fractals*, 138, 2020.
- [12] Li, Lixiang, et al., Propagation analysis and prediction of the COVID-19. *Infectious Disease Modelling*, vol. 5, pp. 282-292, March, 2020.
- [13] Lin Jia, et al., Prediction and analysis of Coronavirus Disease 2019, *Arxiv*, 2020.
- [14] Chintalapudi, Nalini & Battineni, Gopi & Amenta, Francesco. (2020). COVID-19 virus outbreak forecasting of registered and recovered cases after sixty day lockdown in Italy: A data driven model approach. *Journal of microbiology, immunology, and infection*, <https://doi.org/10.1016/j.jmii.2020.04.004>. vol. 53, pp.396-403, April 2020.
- [15] Parbat, Debanjan, and Monisha Chakraborty. "A python based support vector regression model for prediction of COVID19 cases in India." *Chaos, Solitons & Fractals*, 138, 109942, May, 2020.
- [16] Melin, Patricia, et al., Multiple ensemble neural network models with fuzzy response aggregation for predicting COVID-19 time series: the case of Mexico. *Healthcare*. vol. 8. no. 2. Multidisciplinary Digital Publishing Institute, pp. 1-13, June 2020.

- [17] Car, Zlatan, et al., Modeling the Spread of COVID-19 Infection Using a Multilayer Perceptron. *Computational and Mathematical Methods in Medicine*, May, 2020.
- [18] Shastri, Sourabh, et al., Time series forecasting of Covid-19 using deep learning models: India-USA comparative case study, *Chaos, Solitons & Fractals*, 140, 110227, 2020.
- [19] Islam, Md Zahirul, Md Milon Islam, and Amanullah Asraf., A combined deep CNN-LSTM network for the detection of novel coronavirus (COVID-19) using X-ray images. *Informatics in Medicine Unlocked*, 20, 100412, 2020.
- [20] Pal, Ratnabali, et al., Neural network based country wise risk prediction of COVID-19, *arXiv preprint arXiv: 2004.00959*, 2020.
- [21] Liu, Fenglin, et al., Predicting and analyzing the COVID-19 epidemic in China: Based on SEIRD, LSTM and GWR models, *Plos one* 15.8, e0238280, 2020.
- [22] N.P.Dharani et al., Detection of Breast Cancer by Thermal Based Sensors using Multilayered Neural Network Classifier, *International Journal of Engineering and Advanced Technology*, vol. 9 Issue. 2, pp.5615-5618, Dec. 2019.
- [23] C Venkatesh and PolaiiahBojja, Development of Qualitative Model for Detection of Lung Cancer using Optimization, *International Journal of Innovative Technology and Exploring Engineering*, vol. 8 Issue.9, pp. 3143-3147, July 2019.
- [24] C Venkatesh and PolaiiahBojja, A Novel Approach for Lung Lesion Segmentation Using Optimization Technique, *Helix*, vol. 9, no. 1, pp. 4832.
- [25] C Venkatesh and PolaiiahBojja, A Novel Approach for Lung Lesion Segmentation Using Optimization Technique, *Helix*, vol. 9, no. 1, pp. 4832.
- [26] Dharani, N P et al. "Evaluation of Performance of an LR and SVR models to predict COVID-19 Pandemic." *Materials today. Proceedings*, 10.1016/j.matpr.2021.02.166. 16 Feb. 2021, doi:10.1016/j.matpr.2021.02.16.

AUTHORS' PROFILE



Mrs. N.P.Dharani received her M.Tech degree in Electronics and Communication Engineering with specialization of Digital Systems and Computer Electronics from Balaji Institute of Technology and Sciences, Warangal, affiliated to JNTUH, Hyderabad and received B.Tech degree in Electronics and Instrumentation Engineering from Sreenivasa Institute of Technology And Mngement Studies, Chittoor, affiliated to JNTUH, Hyderabad. She is pursuing Ph.D at K L E F Deemed to be University, Vaddeswaram, Guntur, Andhra Pradesh, India. Her area of research is Image/signal Processing, Prediction and Machine Learning/Deep Learning. She has published nine International journals and attended National and International Conferences. She is a life member of International Association of Engineers (IAENG).



POLAIAH BOJJA is presently working as Professor in department of Electronics and communication Engineering, K L E F Deemed to be University, Vaddeswaram, Guntur, Andhra Pradesh, India. He obtained B.Tech from JNTU, Hyderabad. M.Tech from Annamalai University, Tamil Nadu and Ph.D from JNTU Anantapuramu, Andhra Pradesh. He is a principal Investigator for the projects funded by AICTE, New Delhi and DeitY, New Delhi and Collaboration with C-DAC, Thiruvananthapuram. His overall teaching experience is 16 Years. He was published papers in International and National reputed journals with scopus/SCI Indexed and also overall presented conferences in National and International are more than 85. His research interests include Signal Processing and Embedded control systems and also Process Modelling, Optimization, Prediction, Adaptive Optimal Control, Automation, Machine Learning and Computing for Engg., applications. He is a Life Member's of International Society for Technology in Education (ISTE) and (Indian Science Congress Association (ISCA), Member of Automatic Dynamic Controller for systems (ADCS) and International Association of Engineers (IAENG). He has published two text book in Digital image Processing for Industrial Applications and twopatents. He acted as reviewer for many IEEE international conferences and also delivered keynote speech in various conferences and Organizations.

A Proposed Fraud Detection Model based on e-Payments Attributes a Case Study in Egyptian e-Payment Gateway

Mohamed Hassan Nasr¹, Mona Mohamed Nasr³
Faculty of Computers and Information System
Helwan University, Cairo, Egypt

Mohamed Hassan Farrag²
Faculty of Computers and Information System
Fayoum University, Fayoum, Egypt

Abstract—As per Payfort's 2017 report, titled State of payments in the Arab world; Egypt had a 22% yearly increase in the overall volume of internet payments in 2016, which was assessed at \$6.2 billion. e-Payments are the major point of life nowadays in Egypt and the whole world; with tens of e-payments companies in Egypt and more than 5 million transactions done every day and 60 billion EGP volume of payments in 2018. Online and mobile fraud was estimated at \$10.7 billion in 2015, as per Juniper Research, and is expected to reach \$25.6 billion by the end of the decade. As the whole e-payments business is affected by fraud, e-payments firms and their consumers lose a lot of money. On the other hand, one of the most powerful techniques that could be used for fraud predictive is data mining techniques such as the decision tree. This paper introduces a prediction model for managing the risk of fraud in the Egyptian e-payment market that helps to reduce the loss of money. This model is developed using a real dataset from one of Egypt's top e-payment gateways based on the e-payment transaction attributes importance like transaction time, transaction amount, transaction limit, and transaction customer No. repetition limit. The importance of these attributes was determined using IBM SPSS modeler's decision tree and its predictors' importance. The model significantly assisted in the reduction of fraud cases by a very high rate, with an accuracy of 88.45% and a precision of 93.5% resulting in a savings of 101970.52 EGP out of 131297.83 EGP.

Keywords—Data mining; decision tree; e-payments; fraud detection; e-payment gateways; e-commerce

I. INTRODUCTION

Egypt scored the highest growth of online shopping in the Arab world with a 32% increase in the volume of e-payments. According to the internet, world stats report (2017) internet users in Egypt By the end of March 2017, Egypt had officially hit 36.5 percent of the population, half of whom use e-commerce services for everything from purchasing goods and services to paying bills. There are many definitions for e-payments, of which an e-payment system is a form of financial commitment that involves the buyer and the seller facilitated via the use of electronic communications [6]. Another definition defines e-payment as any form of fund transfer via the internet [6].

Using e-commerce; business payments have taken the form of exchanging money electronically and are called electronic payments [2]. Nowadays, most organizations, companies, and

government agencies have adopted electronic commerce to increase their productivity or efficiency in trading products or services in areas such as credit cards, telecommunication, healthcare insurance, automobile insurance, online auction, etc. [1]. The success of a particular electronic payment system is determined by how well it overcomes the practical and analytical hurdles that various online payment methods face.

These challenges include issues of laws and regulations (buyer and seller protection), technological capabilities of e-payment service providers, commercial relationships, and security considerations such as verification and authentication issues [2]. E-commerce systems are used by both legitimate users and fraudsters. Hence, they become more vulnerable to large-scale and systematic fraud. Internet Crime Complaint Centre (IC3) is a valuable resource for both victims of internet crime and law enforcement agencies in identifying, investigating, and prosecuting these crimes. In 2020, the IC3 gathered 15,421 Tech Support Fraud complaints from victims in 60 countries. The losses were over \$146 million, an increase of 171 percent over the previous year (IC3, 2020).

The Egyptian e-payments market is also affected by fraud crimes. Many customers have been defrauded in the Egyptian e-payments market by someone calling them and pretending to be from the e-payments company sales team, attempting to convince them to give their account details in order to receive a bonus. The motivation of this model is to protect and minimize the losses of customers who have been deceived.

One of the most important and rapidly growing payment methods in Egypt is e-payments gateway companies with more than 5 million transactions done every day. Those companies have more than 789 services that customers can use and about 294 thousand outlets spreading in all places in Egypt with 60 billion EGP volume of payments in 2018 and expectation to reach 90 billion in 2019. With high efficiency, ease, and speed, e-payment has become a significant facilitating engine in e-commerce through e-business success.

Hence, the paper is proposing a model for fraud detection in e-payments, especially in Egyptian e-payments companies and the proposed model will be applied to one of those companies. A decision tree, which is one of the most effective data mining techniques, was used to create the model depending on the importance of the e-payment transaction's attributes. The paper is organized into seven sections. The

second section presents definitions of e-payment gateway, decision tree, C5.0 algorithm, and related work. In Section 3 the methodology is presented. Section 4 shows the decision tree model and Section 5 presents results and discussion. Conclusion and future work are presented in Sections 6 and 7.

II. LITERATURE REVIEW

A. Background

This section explains the main points of the area being researched and gives an overview of these points. It starts with explaining what is e-payment gateway then give a brief about decision tree technique, C5.0 algorithm, Splitting criteria, and information gain metric.

1) *e-Payment gateway*: An electronic payment gateway system is a software service that connects with retailer and service provider networks and enables consumers to make payments through these [8]. e-Payment companies offer financial services to consumers and businesses through various channels and a large network of agents. These financial services include paying bills, paying vouchers, reservations, donations, and other services. e-Payment helps businesses to reach more customers, increase productivity, and help consumers to save time and pay for services at any time in a cheaper, easier, faster, and real-time way.

2) *Decision tree*: Decision tree is the most important and widely used categorization and forecasting approach. A decision tree is a tree structure that looks like a flowchart, with each internal node representing an attribute test, each branch reflecting the test's outcome, and each leaf node (terminal node) storing a class label[12]. By learning simple decision rules derived from data properties, a decision tree is used to develop a model that forecasts required variable values. ID3, C4.5, C5.0, and CART are only a few of the algorithms for learning decision trees from a given data set that have been presented. The C5.0 algorithm will be used in our model because of its accuracy and ease of implementation.

3) *C5.0 Algorithm*: One of the most well-known algorithms is C5.0. The C5.0 technique has become the best choice for generating decision trees since it works successfully for most kinds of challenges straight out of the box. The decision trees of the C5.0 algorithm work nearly as well as more difficult and complex machine learning approaches (such as Neural Networks and Support Vector Machines), but are significantly easier to understand and use.

4) *Splitting criteria*: Information Gain (Entropy) is the splitting criterion used by C5.0. The C5.0 model breaks the sample into fields based on whatever field provides the maximum information gain. Each sub-sample specified by the first split is split a second time, generally on a different field, and the technique is repeated until the subsamples cannot be split anymore. Lastly, the lowest-level splits are inspected again, and those that do not add significantly to the model's value are trimmed or removed.

5) *Information gain*: Information gain is a metric for how much data a feature offers about a class. It helps to specify the

order of attributes in the decision tree's nodes. The entropy of the dataset before and after a transformation is used to calculate information gain. The entropy of a sub split can be defined as a measure of its purity. Entropy is always between 0 and 1. Below is how the Information Entropy is calculated for a dataset with N classes.

$$E = - \sum_{i=1}^N p_i \log_2 p_i \quad (1)$$

Where p_i is the probability of randomly picking an element of class i (i.e. the proportion of the dataset made up of class i). Below is the formula of calculating Information Gain based on the calculated Entropy.

$$Gain = E_{parent} - E_{parent} \quad (2)$$

B. Related Work

Techniques for identifying fraud in e-payments are growing very quickly. Some of the popular techniques are rule-based systems, neural networks, Decision trees, machine learning business intelligence, hidden Markov Model, etc. As per lae Chouiekha and EL Hassane Ibn EL Haj [3], they have proposed a model to detect fraudsters in mobile communication using deep learning techniques. In order to forecast fraudulent events in a mobile environment, researchers compared the performance of convolution neural networks to that of classic machine learning algorithms.

Shirley Wong and Sitalakshmi Venkatraman [10] have proposed financial fraud detection and propose a forensic accounting framework using business intelligence. This framework presents a three-phase methodology for performing financial analysis, such as ratio analysis, for a business case scenario using unique knowledge discovery techniques. In contrast to traditional methods of vertical and horizontal analysis for the business case study, the framework's implementation practically demonstrates how the technologies and investigative methods of trend analysis could be used to investigate fraudulent financial reporting using their accounting data.

In another work, Roland Rieke, Maria Zhdanova, Jürgen Repp, and Romain Giot [9] developed a tool for runtime predictive security analysis that analyses process behavior in relation to transactions within a money transfer service and attempts to match it with the expected behavior provided by a process model. The tool analyzes deviations from the given behavior specification for anomalies that indicate a possible misuse of the service related to money laundering activities.

Memorie Mwanza [7] has proposed a model for detection of fraud on tax data for Zambia Revenue Authority using business intelligence model that implements data mining, outlier algorithms for fraud detection and is based on, Continuous Monitoring of Distance-Based and Distance-Based Outlier Queries was then designed and Hao ZHOU, Hong-feng CHAI and Maolin QIU [11] introduce machine learning algorithms to perform fraud detection of bankcard enrollment. They introduce several traditional machine-learning algorithms and finally choose the improved gradient boosting decision tree (GBDT) algorithm software library for use in a real system, namely, XGBoost also Shailesh S. Dhok, G. R.

Bamnote [4] model the sequence of operations in credit card transaction processing using a Hidden Markov Model (HMM) and show how it can be used for the detection of frauds. An HMM is initially trained on a cardholder's regular behavior. An incoming credit card transaction is considered fraudulent if the trained HMM does not accept it with an enough high probability. The hidden Markov Model aids in achieving high fraud coverage while minimizing false alarms.

With [5], two main techniques and using five classifiers, Sahu, Aanchal, G. M. Harshvardhan, and Mahendra Kumar created models to detect credit card fraud transactions. These two techniques are designed to address the problem of data imbalance which helps to detect fraud transaction.

III. METHODOLOGY

This section explains all the steps done to generate the model starting with describing the used data set, its preparation steps, and the used tools that helped build the model. The data set show available service categories and explains transaction details for the e-payment transaction done through the payments gateway.

A. Data Set

The data set contains real-life data of financial transactions of one of the top five e-payments companies in Egypt. Table I shows the services categories that e-payments companies in Egypt provide to their customers. There are governmental services such as electricity; gas and water there are mobile services for recharging and bill payment and many other services such as donation services, airline services DSL services, and many other services that any Egyptian consumer or corporates needs. Table II shows transaction details, some of those details are received by the system when the transaction is processed and some already exist related to the agent who made the transaction. The data set contains about 92718 records divided into two parts. Training data that contains 64903 transactions about (70%) and testing data set that contains 27814 transactions about (30%).

B. Data Preparation

As shown in Fig. 1 data preparation steps start with data gathering. The data was extracted directly and manually from the data source and exported to excel sheets then data cleaning and validation start by removing extraneous data and outliers, filling in missing values, conformed data into a specified pattern sensitive and private data was hidden, and removing error transactions. The next step was discovering and classifying the data. Data were classified and divided into months, weeks, and days for each agent's account based on three attributes (number of transactions, the amount, number of customer's number repeated).

Data analyzing was the final step. The data were divided into three groups based on the volume of transactions and the volume of their amounts in order to obtain better and more accurate results.

- Low volume rate group.
- Medium volume rate group.
- High volume rate group.

TABLE I. SERVICE CATEGORY

#	Service category	Description
1	Mobile recharge	Used for recharging mobile
2	Mobile bill payment	Paying for mobile monthly bills
3	Mobile e-voucher	Generate vouchers for recharging mobile phone
4	Donations	Donation to charities
5	Airlines	Reserving airlines tickets
6	Gas	Gas bills
7	Water	Water bills
8	Electricity	Electricity bills
9	DSL	DSL bills
10	Cinema tickets	Reserving tickets online
11	Games	Paying for online games

TABLE II. TRANSACTION DETAILS

#	Attribute	Description
1	Create Date/Time	Date and time for creating the transaction
2	Update Date/Time	Date and time for receiving a response from the service provider
3	Transaction ID	Unique transaction ID in the system
4	Provider	Service provider name
5	Service	Service name
6	Customer number	Identifier used by customer EX: Phone number for recharging mobile
7	Amount	The amount that should be paid
8	Total amount	The amount that should be paid + Service charge
9	Status	Status of transaction success or error
10	Provider Response Code	Code sent by the service provider in the response
11	Provider Transaction ID	Service provider unique transaction Id in
12	Transaction Initiator	Name of agent that made the transaction
13	Transaction Deduction From	Account number for the agent who made the transaction
14	Interface	Type of interface he used (mobile, web
15	Outlet	Name of the outlet
16	Area	Area of the outlet

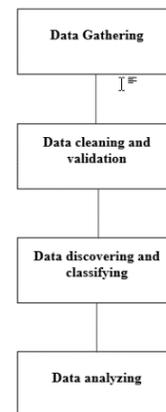


Fig. 1. Data Set Preparation Steps.

TABLE III. DATA CATEGORIES SPECIFICATION

		Low Volume Rate	Medium Volume Rate	High Volume Rate
Transaction	Daily	50	100	300
	Weekly	300	1000	3000
	Monthly	1200	5000	10000
Amount	Daily	1200	2500	8000
	Weekly	5000	20000	80000
	Monthly	20000	100000	500000
Customer No.	Daily	3	6	10
	Weekly	15	30	50
	Monthly	50	100	100

As shown in Table III, each group contains the maximum daily transactions number allowed for the agent, the maximum weekly number, the maximum monthly number also the maximum daily, weekly and monthly amount, and the same for customer No. repeated times.

The data set was divided into six data sets based on the three groups that were received from the data analysis phase. Each group has two data sets one for training and one for testing. The training data set contains 70% of transactions and the testing data set contains 30% of transactions for each group as shown in Table IV.

TABLE IV. DATA SET VOLUME

	Low Volume Rate	Medium Volume Rate	High Volume Rate	Total
Training	42426	28242	22050	64903
Testing	12727	8472	6615	27814

IV. DECISION TREE MODEL

A. Tools used

IBM SPSS modeler was used to create the decision tree model. It is a data-mining tool that allows you to create

predictive models without programming. It helps you to specify groups, identify correlations between them and predicting events that will happen in the future. ASC5.0 is one of the decision tree algorithms included in IBM modeler, and it was used in the proposed model as shown in Fig. 2 due to its efficiency, as stated previously.

B. Splitting Criteria

As previously stated in the background section, C5.0 uses Information Gain (Entropy) as its splitting criteria. Fig. 3, Fig. 4, and Fig. 5 show the predictor importance for the e-payment transaction attributes for each group (low, medium, and high) as generated by IBM modeler. e-Payment transaction attributes are transaction allowed time, transaction (daily, weekly, monthly) allowed limit, transaction (daily, weekly, monthly) allowed amount, and transaction (daily, weekly, monthly) allowed customer number limit.

C. Resulting Decision Tree

The model will start by checking the transaction time if it is in the agent’s allowed time or not and if it is in the allowed time the model will go to the next step to check if the transaction exceeds the maximum monthly transaction number if it did not exceed the model will check for the weekly and after that the daily. The model will check for Amount and customer, No. repeated in the same way as the transaction number and if the transaction passed all the conditions it will be accepted. Fig. 6 shows the whole procedure of the resulting decision tree.

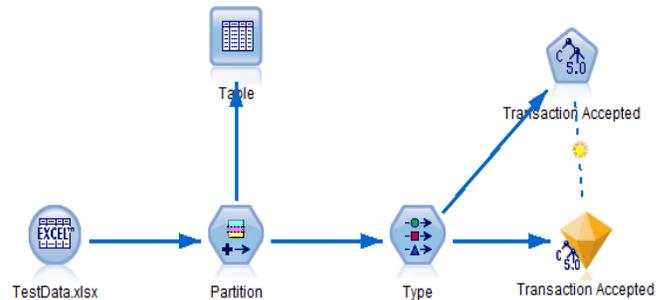


Fig. 2. IBM SPSS Decision Tree.

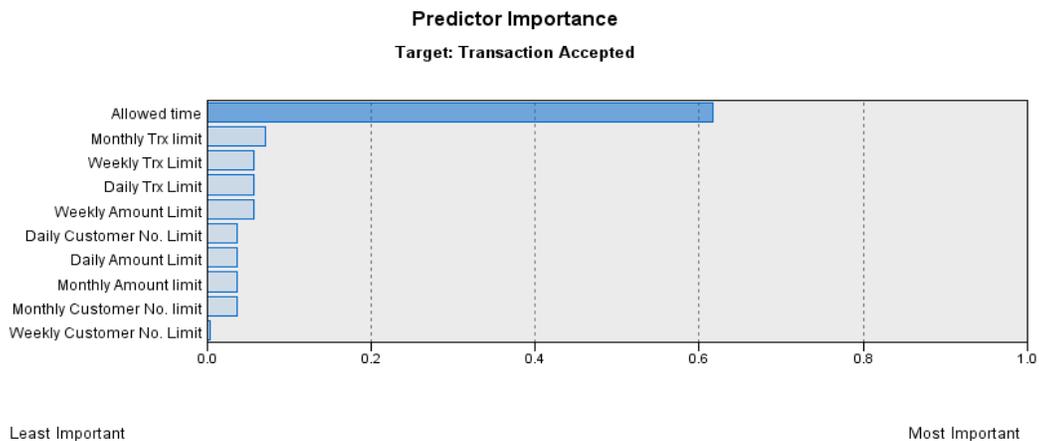


Fig. 3. Predictor Importance for Low Rate Group.

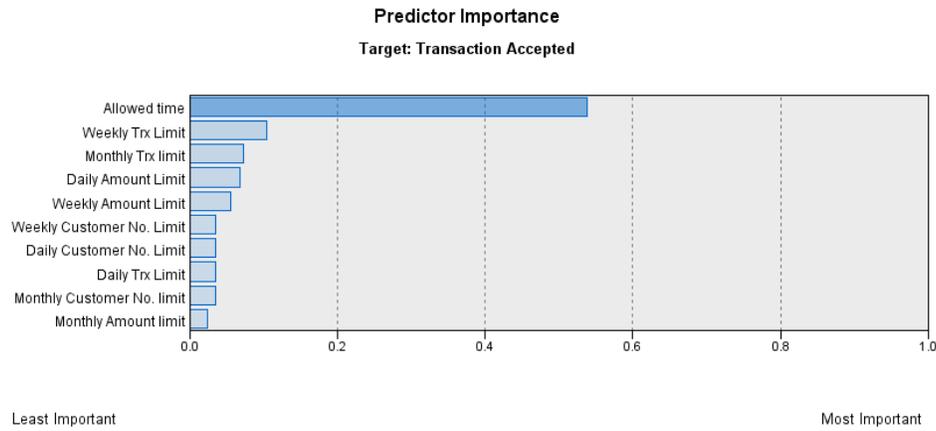


Fig. 4. Predictor Importance for Medium Rate Group.

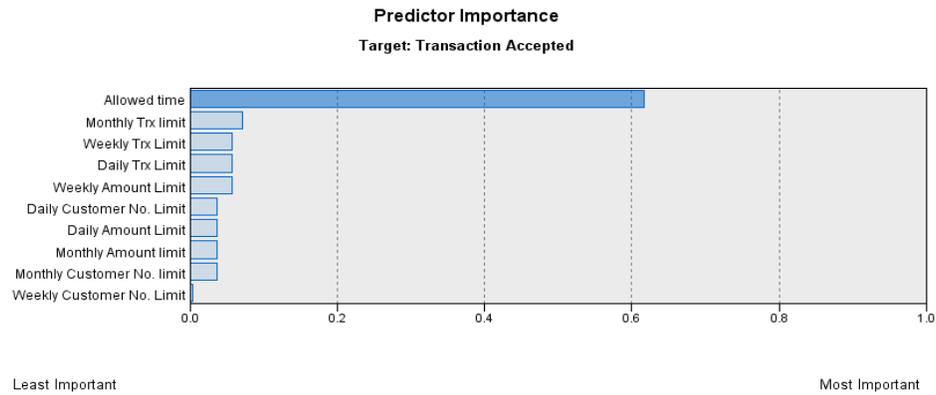


Fig. 5. Predictor Importance for High Rate Group.

D. Decision Tree Derived Rules

R = Result for condition statement.

1 = True, 0 = False.

F(Mtrx) = Monthly transaction.

F(Mamut) = Monthly Amount.

F(Mcus.no) = Monthly Customer No.

F(Wtrx) = Weekly transaction.

F(Wamut) = Weekly Amount.

F(Wcus.no) = Weekly Customer No.

F(Dtrx) = Daily transaction.

F(Damut) = Daily Amount.

F(Dcus.no) = Daily Customer No.

1) Low volume rate group

$$R = f(Mtrx) \int_{0, > 1200}^{1, \leq 1200} R = f(Wtrx) \int_{0, > 300}^{1, \leq 300} R = f(Dtrx) \int_{0, > 50}^{1, \leq 50} \quad (3)$$

$$R = f(Mamut) \int_{0, > 20000}^{1, \leq 20000} R = f(Wamut) \int_{0, > 5000}^{1, \leq 5000} R = f(Damut) \int_{0, > 1200}^{1, \leq 1200} \quad (4)$$

$$R = f(Mcus.no) \int_{0, > 50}^{1, \leq 50} R = f(Wcus.no) \int_{0, > 15}^{1, \leq 15} R = f(Dcus.no) \int_{0, > 3}^{1, \leq 3} \quad (5)$$

2) Medium volume rate group

$$R = f(Mtrx) \int_{0, > 5000}^{1, \leq 5000} R = f(Wtrx) \int_{0, > 1000}^{1, \leq 1000} R = f(Dtrx) \int_{0, > 100}^{1, \leq 100} \quad (6)$$

$$R = f(Mamut) \int_{0, > 100000}^{1, \leq 100000} R = f(Wamut) \int_{0, > 20000}^{1, \leq 20000} R = f(Damut) \int_{0, > 2500}^{1, \leq 2500} \quad (7)$$

$$R = f(Mcus.no) \int_{0, > 100}^{1, \leq 100} R = f(Wcus.no) \int_{0, > 30}^{1, \leq 30} R = f(Dcus.no) \int_{0, > 6}^{1, \leq 6} \quad (8)$$

3) High volume rate group

$$R = f(Mtrx) \int_{0, > 10000}^{1, \leq 10000} R = f(Wtrx) \int_{0, > 3000}^{1, \leq 3000} R = f(Dtrx) \int_{0, > 300}^{1, \leq 300} \quad (9)$$

$$R = f(Mtrx) \int_{0, > 500000}^{1, \leq 500000} R = f(Wtrx) \int_{0, > 80000}^{1, \leq 80000} R = f(Dtrx) \int_{0, > 8000}^{1, \leq 8000} \quad (10)$$

$$R = f(Mtrx) \int_{0, > 100}^{1, \leq 100} R = f(Wtrx) \int_{0, > 50}^{1, \leq 50} R = f(Dtrx) \int_{0, > 10}^{1, \leq 10} \quad (11)$$

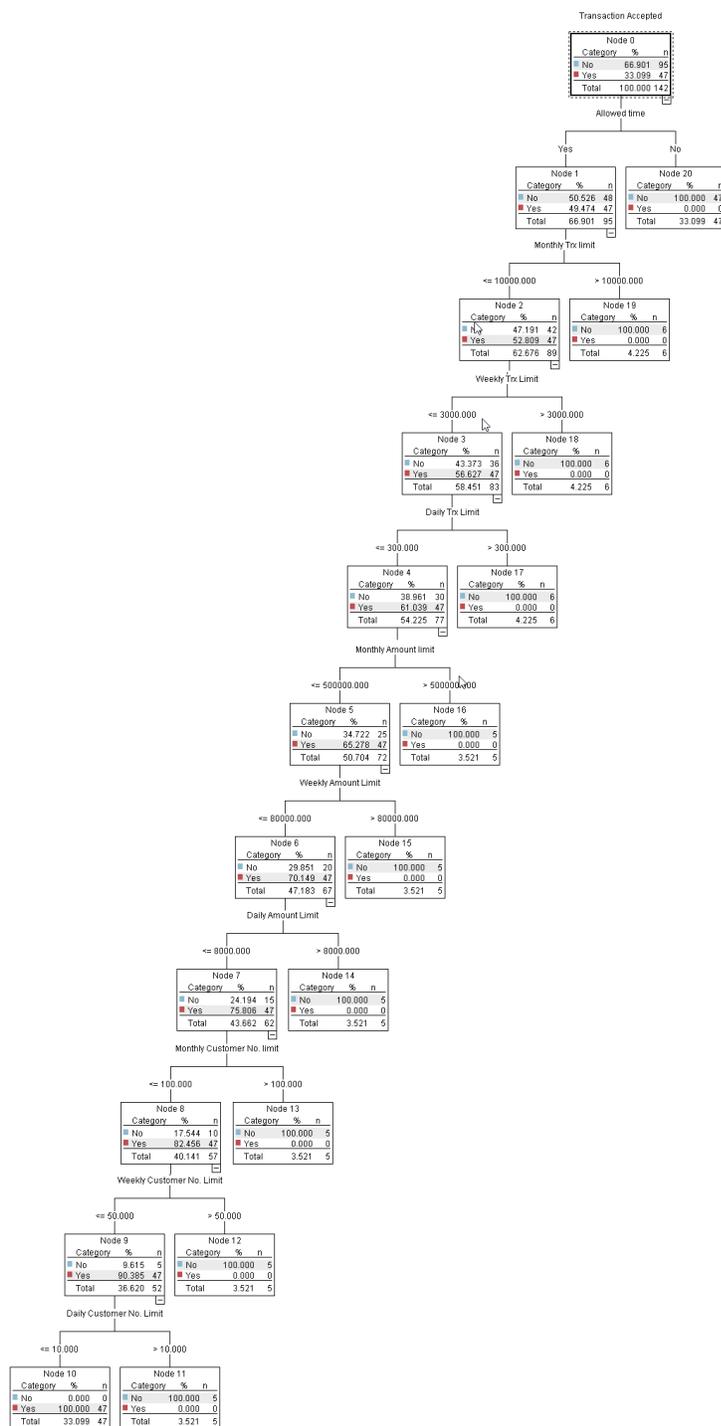


Fig. 6. Decision Tree Model for Low, Medium and High Volume Rate Groups.

V. RESULT AND DISCUSSION

A. Results and Findings

The model was implemented and a real data set that containing fraud transactions were used to test the implemented model. The used data set is a real dataset of a total of 1590 transactions with 590 fraud transactions totaling 131297.83 EGP in losses. With a total of 433 out of 590 fraud

transactions detected, the model was able to identify more than 73%. Out of a total of 131297.83 EGP, the model saved 77% of the total amount lost due to fraud, equating to 101970.52 EGP. The model significantly reduced fraud transactions, demonstrating the need to analyze and verify that all customer transactions are carried out by them, not only that they come from their account. The confusion matrix was used to calculate the accuracy, precision, and false alarm rate for the model as shown in Table V.

TABLE V. CONFUSION MATRIX DEFINITIONS

TP	True positive (number of transactions that were fraudulent and were also classified as fraudulent by the model)
TN	True negative (number of transactions that were legitimate and were also classified as legitimate)
FP	False positive (number of transactions that were legitimate but were wrongly classified as fraudulent transactions)
FN	F False negative (number of transactions that were fraudulent but were wrongly classified as legitimate transactions by the model)

Accuracy is the fraction of transactions that were correctly classified.

$$\text{Accuracy (ACC)/Detection rate} = (TN + TP) / (TP + FP + FN + TN)$$

Precision (also known as the detection rate), the number of transactions either genuine or fraudulent were correctly classified.

$$\text{Precision/Detection rate/Hit rate} = TP / TP + FP$$

False Alarm rate measures out of total instances classified as fraudulent or how many were wrongly classified.

$$\text{False Alarm Rate} = FP/FP+TN.$$

$$TP= 433 \quad TN= 1000 \quad FP= 30 \quad FN= 157.$$

$$\text{Accuracy (ACC)/Detection rate} = 1000 + 433 / (433 + 30 + 157 + 1000) = 88.45 \%$$

$$\text{Precision/Detection rate/Hit rate} = 433/433 + 30 = 93.5 \%F.$$

$$\text{False Alarm Rate} = 30/30+1000 = 2.9\%VI. Accuracy \& \text{ Precision \& False Alarm Rate.}$$

Accuracy	Precision	False Alarm Rate
88.45 %	93.5 %	2.9 %

B. Limitation of Work

One of the limitations that we faced is the lack of previous research studies on the topic of the Egyptian e-payment market and the absence of official statistics that help estimate the extent of the problem. As well as the difficulties we faced as a result of the payment companies' lack of cooperation and refusal to provide any numbers of fraud losses suffered by their customers.

VI. CONCLUSION

The paper introduces a fraud detection model using data mining. The model used the decision tree technique. The model is based on real data from one of the Egyptian e-payment gateways. With an accuracy of 88.45 percent, the suggested model helps in the detection of any up normal transactions that differ from the typical behavior of users' transactions. Fig. 7 shows that the model detected 433 fraud transactions out of 590, resulting in a savings of 101970.52 EGP out of 131297.83 EGP as shown in Fig. 8. Using this model, a secure payment environment will enhance user confidence and reduce money loss.

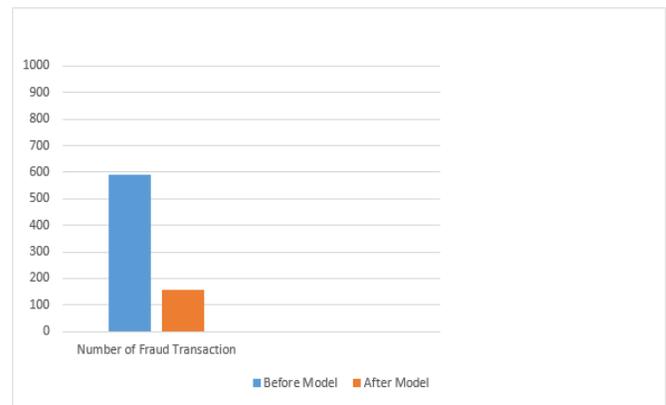


Fig. 7. Number of Fraud Transactions before and after using the Model.

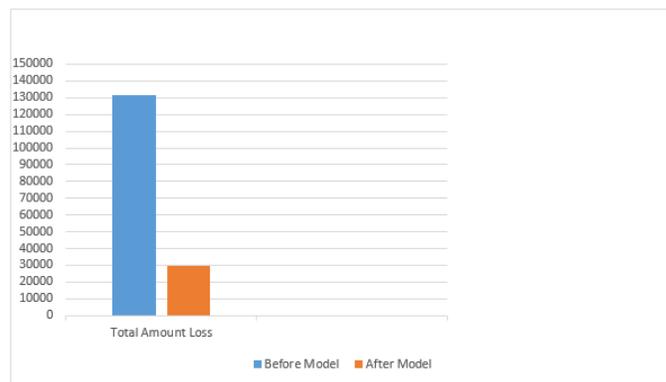


Fig. 8. Total Amount Loss before and after using the Model.

VII. FUTURE WORK

This paper introduced a fraud detection model using the decision tree technique for a specific type of e-payment, which are e-payment companies that allow users to pay for services through their system. Through this research, several points have arisen that must be discussed in the future. One of these points is to implement the proposed model to other different e-payment types and analyze the results in order to improve the Another point is to apply this model using other data mining techniques and compare it to decision tree technique also applying this model in another e-payments fields for example credit card payments and mobile banking payments.

REFERENCES

- [1] Abdallah, Aisha, Mohd Aizaini Maarof, and Anazida Zainal. "Fraud detection system: A survey." Journal of Network and Computer Applications 68 (2016): 90-113.
- [2] Bezovski, Zlatko. "The future of the mobile payment as electronic payment system." European Journal of Business and Management 8, no. 8 (2016): 127-132.
- [3] Chouiekh, Alae, and EL Hassane Ibn EL Haj. "Convnets for fraud detection analysis." Procedia Computer Science 127 (2018): 133-138.
- [4] Dhok, Shailesh S., and G. R. Bamnote. "Credit card fraud detection using hidden Markov model." International Journal of Soft Computing and Engineering (IJSC) 2, no. 1 (2012): 231-237.

- [5] Sahu, Aanchal, G. M. Harshvardhan, and Mahendra Kumar Gourisaria. "A dual approach for credit card fraud detection using neural network and data mining techniques." In 2020 IEEE 17th India Council International Conference (INDICON), pp. 1-7. IEEE, 2020.
- [6] Kabir, Mohammad Auwal, Siti Zabedah Saidin, and Aidi Ahmi. "Adoption of e-payment systems: a review of literature." In International Conference on E-Commerce, pp. 112-120. 2015.
- [7] Mwanza, Memorie. "Fraud detection on big tax data using business intelligence, data mining tool: A case of Zambia revenue authority." PhD diss., University of Zambia, 2017.
- [8] Nasr, Mohamed Hassan, Mohamed Hassan Farrag, and Mona Nasr. "e-payment Systems Risks, Opportunities, and Challenges for Improved Results in e-business." International Journal of Intelligent Computing and Information Sciences 20, no. 1 (2020): 16-27.
- [9] Rieke, Roland, Maria Zhdanova, Jürgen Repp, Romain Giot, and Chrystel Gaber. "Fraud detection in mobile payments utilizing process behavior analysis." In 2013 International Conference on Availability, Reliability and Security, pp. 662-669. IEEE, 2013.
- [10] Wong, Shirley, and Sitalakshmi Venkatraman. "Financial accounting fraud detection using business intelligence." Asian Economic and Financial Review 5, no. 11 (2015): 1187-1207.
- [11] Zhou, Hao, Hong-feng Chai, and Mao-lin Qiu. "Fraud detection within bankcard enrollment on mobile device based payment using machine learning." Frontiers of Information Technology & Electronic Engineering 19, no. 12 (2018): 1537-1545.
- [12] Nuruzzaman, Md, Md Shahadat Hossain, Md Mostafijur Rahman, Ahete Shamul Haque Chowdhury Shoumik, Md Abbas Ali Khan, and Md Tarek Habib. "Machine Vision Based Potato Species Recognition." In 2021 5th International Conference on Intelligent Computing and Control Systems (ICICCS), pp. 1-8. IEEE, 2021.

Improved ISODATA Clustering Method with Parameter Estimation based on Genetic Algorithm

Kohei Arai

Graduate School of Science and Engineering
Saga University, Saga City
Japan

Abstract—Improved ISODATA clustering method with merge and split parameters as well as initial cluster center determination with GA: Genetic Algorithm is proposed. Although ISODATA method is well-known clustering method, there is a problem that the iteration and clustering result is strongly depending on the initial parameters, especially the threshold for merge and split. Furthermore, it shows a relatively poor clustering performance in the case that the probability density function of data in concern cannot be expressed with convex function. To overcome this situation, GA is introduced for the determination of initial cluster center as well as the threshold of merge and split between constructing clusters. Through experiments with simulated data, the well-known the University of California, Irvine: UCI repository data for clustering performance evaluations and ASTER/VNIR: Advanced Spaceborne Thermal Emission and Reflection Radiometer / Visible and Near Infrared Radiometer onboard Terra satellite of imagery data, the proposed method is confirmed to be superior to the conventional ISODATA method.

Keywords—ISODATA clustering; nonlinear merge and split; concaveness of probability density function: PDF; remote sensing satellite imagery data; clustering; genetic algorithm: GA; nonlinear optimization

I. INTRODUCTION

Clustering methods can be broadly divided into hierarchical and non-hierarchical clustering methods [1], [2]. Typical non-hierarchical clustering methods are the k-means method¹ and the ISODATA method². In the k-means method, it is necessary to give the number of clusters and the initial cluster center in advance, and the calculation time and the obtained cluster shape will change depending on their settings. On the other hand, the ISODATA method can autonomously determine the effective number of clusters within a certain range of the set number of clusters.

The ISODATA method separates target data individuals into clusters based on the k-means method, and then divides and fuses the clusters according to a preset threshold value based on statistical indices within and between clusters, and rearranges the cluster individuals. This is a method of repeating this series of processing until the rearrangement end judgment criterion is satisfied. Therefore, although the ISODATA method has relatively high clustering accuracy, it takes a considerable amount of processing time to set these

parameters, and this determination can often be difficult. The final number of clusters, clustering accuracy and calculation time depends on these parameters. Especially in the case of clustering for multi-dimensional data such as satellite images, the number of parameters increases as the number of dimensions increases, so adjustment is extremely difficult.

Moreover, the k-means method and the ISODATA method implicitly assume that the probability density function of the target data is a convex function. That is, high clustering accuracy cannot be expected in the case of a distribution of a cluster that is a concave function in a multidimensional space.

This paper assumes that the target data is multidimensional like a satellite image, deals with the case where the probability density function of the cluster individual is concave, and improves the clustering accuracy by using the estimated optimal parameters. The author proposes a clustering method based on the ISODATA method. Genetic Algorithms (GA) [3], [4] were used for parameter optimization. There is the other alternative algorithm, sand cat swarm optimization (SCSO) [5], Grey Wolf Optimizer (GWO) algorithm [6], Moth-flame optimization algorithm [7], etc.

GA does not guarantee a global optimal solution, but it is a probabilistic optimization method that can estimate a suboptimal solution in a relatively short time [8], [9], [10]. It is effective for problems that have not been discovered and have such a large solution space that a full search is considered impossible [11], [12]. In this paper, the author first shows the effectiveness of the proposed method using simulation data in which the distribution of cluster individuals is a concave function.

The author also applied the UCI repository data set [13] and satellite image data, which are frequently used for comparative evaluation of clustering accuracy, to the proposed method and evaluated the clustering accuracy. The conventional ISODATA method and shape-independent clustering were used as the conventional clustering methods [14], [15]. It is reported here because the proposed method was superior to the conventional methods.

The following section describes related research works. Then the proposed method is described followed by experiment. After that, conclusion is described together with some discussions.

¹ https://en.wikipedia.org/wiki/K-means_clustering

² <https://www.harrisgeospatial.com/docs/ISODATAClassification.html>

II. RELATED RESEARCH WORK

Influence due to geomorphology on context Genetic Algorithm: GA clustering is investigated [16]. On the other hand, learning processes of image clustering method with density maps derived from Self-Organizing Mapping (SOM) is also proposed [17].

Non-linear merge and split method for image clustering (closely related to the proposed method here) is proposed [18]. Meanwhile, revised pattern of moving variance for acceleration of automatic clustering is investigated [19].

Automatic detection method for clustered micro calcification in mammogram image based on statistical texture features is proposed [20]. On the other hand, comparative study between the proposed GA based ISODATA clustering and the conventional clustering methods are conducted [21].

Image clustering method based on density maps derived from Self Organizing Mapping: SOM is proposed [22]. Meanwhile, clustering method based on Messy Genetic Algorithm: GA for remote sensing satellite image clustering is proposed [23].

Visualization of learning process for back propagation Neural Network: NN clustering is proposed [24]. On the other hand, improvement of automated detection method for clustered micro calcification base on wavelet transformation and support vector machine is attempted [25].

Image clustering method based on Self-Organization Mapping: SOM derived density maps and its application for Landsat Thematic Mapper: TM image clustering is proposed [26]. Also, comparative study between the proposed shape independent clustering method and the conventional method (k-means and the others) is conducted [27].

Genetic Algorithm: GA utilizing image clustering with merge and split processes which allows minimizing Fisher distance between clusters is proposed [28]. Also, Fisher distance-based GA clustering taking into account overlapped space among probability density functions of clusters in feature space is proposed [29].

Initial centroid designation algorithm for k-means clustering is proposed [30]. Meanwhile, Pursuit Reinforcement Competitive Learning: PRCL based online clustering with tracking algorithm and its application to image retrieval is proposed [31].

III. PROPOSED METHOD

Clustering is a method of classifying target data by collecting similar items based on the similarity or dissimilarity between target data individuals and generating groups (clusters); also called analysis. The criteria for measuring how similar the target data individuals are similarity (dissimilarity) and dissimilarity. The degree of similarity is, for example, a measure indicating that the object is more similar as the value is larger, such as the correlation coefficient.

On the other hand, the dissimilarity, which is also called the dissociation degree, is a measure indicating that the larger the value, the less similar the objects. Generally, dissimilarity (distance) is often used. The similarity defined in clustering

has indices such as matching coefficient and similarity ratio, while there are many definitions of distance.

In this paper, we use local mean distance and similarity that can deal with the case where the probability density function of the target data individual is a concave function, and it is necessary for the ISODATA method by GA using the fitness function based on these. We propose a clustering method that determines the initial cluster centers as well as the thresholds for cluster partitioning and fusion.

A. Local Average Distance and Local Similarity

The author proposes the Moving Window method. Once the local range (Window) is determined, the local inter-individual distance and the similarity within the range are obtained. The Moving Window method is a method of finding the sum or average of local distances and similarities over the entire area by moving the local area little by little along each dimension (Fig. 1). In this research, the author uses a hypersphere as a window in multidimensional Euclidean space. In the figure, while moving in the local range of radius r , the distance between individuals within this range, or the sum and average of the similarities are obtained.

As shown in Eq. (1), the average value of the inter-individual distances within the local range is called the local average distance L .

$$L = \frac{1}{n(n-1)} \sum_{i=1}^n \sum_{j=1, j \neq i}^n \delta_{i,j} \|l_i - l_j\| \quad (1)$$

where,

$$\delta_{i,j} = \begin{cases} 1, & \text{if } C_i = C_j \\ \frac{1}{2}, & \text{if } C_i \neq C_j \end{cases} \quad (2)$$

In addition, n is the number of individuals in the local range, l is the position vector of the individual, and C_k is the membership cluster of the individual of k . That is, the distance (norm) between individuals is the sum of all the distances between individuals in the local range considering 1 for the same cluster and half the weight for different classes.

Sum of local similarity is defined as follows:

The difference in distance between individuals belonging to the same cluster as the diameter within a certain local range is taken as the similarity, and the sum DS thereof is obtained as in the following formula.

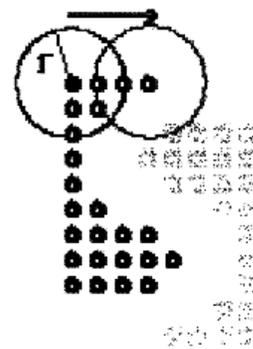


Fig. 1. Moving Window Method.

$$DS = \sum_{i=1}^n \sum_{j=1, j \neq i}^n \delta_{i,j} (2r - ||l_i - l_j||) \quad (3)$$

where,

$$\delta_{i,j} = \begin{cases} 1, & \text{if } C_i = C_j \\ 0, & \text{if } C_i \neq C_j \end{cases} \quad (4)$$

B. Determine the Size and Weight of the Local Range

The local mean distance has a role to detect inter-cluster variance. Therefore, it is desirable that the local range, that is, the window size in the Moving Window method, can cover the inter-cluster dispersion. When the window size, that is, the radius of the hypersphere is gradually increased from the minimum distance between individuals, and the total sum of the local average distances is obtained, the local average distance has a peak when the inter-cluster dispersion is covered.

This peak is found and the radius of the hypersphere in its local range is used as the reference r_p . Experiments were performed using the simulation data shown in Fig. 2(a), (b), and (c) while changing the radius of the hypersphere. The results are summarized in Table I.

Looking at Table I, even if the radius of the hypersphere is made larger than the standard to some extent, it does not affect the clustering result, but below the standard value, it greatly affects the cluster result.

This can identify both cluster individuals if r is set so as to include the maximum inter-individual distance between different clusters, but if set shorter than this, that is, if the local range is set narrow, both cluster individuals are separated. It means that it will be difficult. Therefore, it is sufficient to set γ sufficiently long, but this is directly related to the increase in processing time (proportional to r^2), which is a problem.

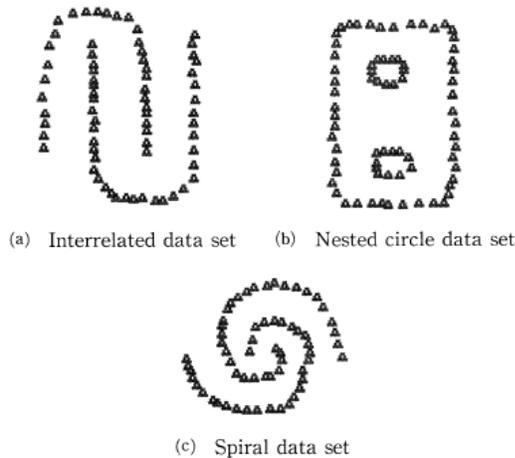


Fig. 2. 2D Simulation Data.

TABLE I. EFFECT OF WINDOW SIZE ON LOCAL AVERAGE DISTANCE

Error(%)	0.5 r_p	1.0 r_p	1.5 r_p	2.0 r_p
Interrelated	98.55	0.00	0.00	0.00
Nested circle	97.30	0.00	0.00	0.00
Spiral	96.77	96.77	0.00	0.00

Therefore, in this study, twice the reference value, $2r$, was taken as the radius of the hypersphere. On the other hand, the sum of local similarities is related to the distance between individuals within a cluster. The window size should be longer than the distance between all individuals in the cluster and their nearest neighbors, and smaller than the inter-cluster variance. When the window size is small, the clustering result has a large number of clusters due to the local average distance, and in an extreme case, all individuals themselves become clusters.

Therefore, the validity of the window size can be judged from the final number of clusters. Assuming that there is no isolated individual, it is appropriate that the window size, that is, the diameter of the hypersphere, is longer than the distance (d_{min}) between the individual farthest from other individuals and their nearest neighbors. First, clustering is performed by setting the radius of the hypersphere to d_{min} . Clustering is performed by increasing the radius of the hypersphere when the number of final clusters exceeds twice the expected number, or when isolated individuals appear. This process was repeated until the final number of clusters fell within twice the expected number.

It is important for the goodness of fit so that individuals whose distance is less than a certain threshold value belong to one cluster, that is, the sum DS of local similarities is maximized. This is equivalent to making the local average distance L as small as possible. From this point of view, it is better to make the weight M smaller, but the effective number of DS is finite, so the value of the weight must be set so that the effective value of M/L can influence the effective number of DS . Also, since maximizing DS is prioritized, it must be $M/L \ll DS$.

C. GA

1) *Real-Coded Genetic Algorithms (RCGA)*: GA is an optimization algorithm that refers to the evolution of living organisms. In GA, the solution of the problem is expressed as an individual, and each individual is composed of chromosomes. Individuals evolve by selective selection, crossover, and mutation, and search for optimal solutions. Early GAs performed crossovers and mutations with binary coded bit strings of variables. This ignores the continuity of variables.

On the other hand, a GA has been proposed which uses the numerical value itself and performs crossover and mutation considering the continuity of variables. This is called Real-Coded Genetic Algorithms (RCGA). RCGA does not use the bit string like the conventional GA but expresses the individual used for the search by converting it into a real vector.

Therefore, the conventional GA chromosomal locus composition replaces the bit string with a real vector. At this time, the multidimensional vector centered on the initial cluster is used as the real vector. In this research, the initial cluster center optimization is performed based on RCGA.

2) *Fitness function*: The author defines a fitness function that maximizes the ratio of the sum of inter-cluster variance to the sum of intra-cluster variance. As for the clustering result, the cluster is configured to be optimal when the probability density function of the target data individual can be regarded as a convex function. On the other hand, if the fitness functions are defined as in Eq. (5) using the sum of local similarities and the local average distance, an optimal cluster configuration is possible even in the case of a concave function distribution.

$$F = \frac{\sum DS + M}{\sum L} \quad (5)$$

where F is the goodness of fit and M is the weight. The local mean distance represents the intra-cluster variance in the local range. The smaller this value, the smaller is the intra-cluster variance. If all individuals in the local range belong to individual clusters, the local mean distance is the minimum.

When this value is used as the goodness of fit, the individuals within the range tend to belong to different clusters, and it is possible to perform clustering so that the individuals at both ends of the local area with large inter-individual variance belong to different clusters.

The sum of local similarities is related to the intra-cluster variance in the local range. The larger this value, the smaller the variance within the cluster in the local range. When all individuals in the local range belong to one cluster, the sum of local similarities becomes maximum. When this value is used as the goodness of fit, the number of clusters tends to decrease, and it is possible to perform clustering so that individuals with close distances belong to one cluster.

By using Eq. (5), it is possible to generate a fitness function that does not depend on the shape of the cluster by balancing the sum of the local average distance and the local similarity and considering the intra-cluster variance and inter-cluster variance at the same time.

3) *Crossover*: Since RCGA does not code, a crossover operator specialized for this is required. Typical examples of this include blend crossover (BLX- α) and unimodal normal distribution crossover (UNDX). BLX- α determines offspring as follows.

a) Let α , b be the two parents in the real-valued vector space.

b) Calculate the intersection $[A, B]$.

$$A = \min(a, b) - \alpha |a - b| \quad (6)$$

$$B = \max(a, b) + \alpha |a - b| \quad (7)$$

where α is a coefficient parameter.

c) Determine the offspring from the interval $[A, B]$ with uniform random numbers.

That is, in the conventional GA, unlike the crossover method in which some or all of the loci of chromosomes are replaced, the vector obtained by subtracting the coefficient times the inter-individual distance from the smaller individual vector to the larger individual vector is added. It is a method

of determining a crossing vector according to a uniform random number generated in the range up to it.

4) *Mutation method*: Mutations were given by normal distribution. That is, it is a method of changing the real number vector of an individual according to the normal distribution according to the probability of the mutation to be set.

5) *Selection method*: The selection method of RCGA is the tournament method, and the elite strategy to enhance the convergence performance of RCGA is adopted.

6) *RCGA Specific method and parameter setting*: RCGA is more efficient because there are many real-valued parameters in ISODATA method. In this study, RCGA is used to optimize the thresholds of initial cluster centers and division and fusion in the ISODATA method. The specific method is shown below.

- Set the number of RCGA populations to 30 and the number of generations to 300.
- Since RCGA is used, real coding is constructed by constructing a multidimensional vector from real values of each parameter and performing crossover and mutation on the vector.

The composition of the vector of chromosomes is as follows,

$$C(S, M, A_1, A_2, \dots, A_m) \quad (8)$$

where S is the division threshold, M is the fusion threshold, and A_m is the m -th coordinate of the initial cluster center.

In this way, the chromosome is defined by Eq. (8), and RCGA is performed to find the optimal thresholds for division and fusion of the initial cluster center and ISODATA.

- Set the tournament selection size to 3.
- Use the BLX- α method as the crossover method. The value of α is set to 0.5 and the crossover probability is set to 70%, in order to prevent falling into a local solution and also considering the speed of convergence.
- The mutation method uses the normal distribution mutation method and sets σ to 0.5. Here, if the mutation probability is set to 5% or less, the local solution is likely to fall, and if the mutation probability is increased, the efficiency of GA is deteriorated, so it is set to 5%.

In order to compare the clustering results, we perform clustering of 100 sets of initial cluster centers randomly determined by uniform random numbers and calculate the maximum likelihood result and average value. The author also does some hierarchical clustering.

D. Improved ISODATA

As mentioned above, the ISODATA method is a method in which target data individuals are divided by constructing cluster boundaries by the hyperplane (Voronoi division³) by

³ https://en.wikipedia.org/wiki/Voronoi_diagram

the k-means method. That is, the probability density function of the cluster is implicitly assumed to be a convex function, and if the probability density function of the target data individual is a concave function, accurate cluster division cannot be expected. When the probability density function is concave, the ISODATA method may be able to deal with the division and fusion process, but there is a possibility that clusters that are correctly classified by division and fusion once will be destroyed by rearrangement.

E. Reduction of Calculation Amount

In the proposed method, the initial cluster center is estimated in advance by the real-valued GA (RCGA), so that the clustering result can obtain an almost optimal cluster result without relying on the modification of the cluster center by repeating the ISODATA method. That is, the clustering result is hardly affected even if the number of iterations of the ISODATA method is reduced to some extent. In the experiment, the number of repetitions of the ISODATA method was set to 4.

In the Moving Window method, the computational complexity increases exponentially as the number of dimensions increases. In the proposed method, we decided to reduce the amount of calculation by moving each individual in order, rather than moving gradually along each dimension.

IV. EXPERIMENTS

A. Dataset of Data used

The UCI repository dataset is a data archive published by knowledge discovery researchers at the University of California, Irvine (University of California, Irvine). It can be accessed by anyone on the web page⁴. In this research, Iris, Wine, New thyroid, and Ruspini dataset of R, and fossil dataset of Chernoff are used for the experiment.

1) *Iris data set* The Iris data set is data of three types of iris flowers. The total number of individuals in the dataset is 150, including 50 individuals for each type. The data in this dataset is four-dimensional. They are the width and length of sepal, the width and length of petal, and the unit is cm. The Iris dataset is one of the best-known datasets for clustering.

2) *Wine dataset*: The Wine Dataset is the chemical analysis data for three Italian wines. The total number of individuals in the dataset is 178, and the number of individuals of each type is 59, 71, 48, respectively. The data in this dataset is 13 dimensions. There is a large difference in the range of data in each dimension. There are dimensions where all numbers are less than or equal to 1 and dimensions that include up to 1,000. Therefore, clustering Wine datasets is considered difficult. In this study, the normalized (Min: 0, Max: 1) Wine dataset was used.

3) *Ruspini data set*: The Ruspini Data Set is included in R and S-plus. It is four-dimensional data with four categories. The total number of individuals is 75, including 23, 20, 17, and 15 individuals.

4) *New thyroid data set*: The New thyroid data set is 5D data on infectious diseases in UCI. The total population is 215. The number of categories is 3, category 1 includes 150 individuals, category 2 includes 35, and category 3 includes 30.

5) *Fossil data set*: The fossil data set is 6-dimensional data on three types of limestone by Chernoff. The total number of individuals is 87, and each category includes 40,34,13 individuals.

These five datasets are standard datasets that are often used to compare clustering methods. Experiments were performed on these data sets. The results are summarized in Table II.

It can be seen from Table II that the error is 20.23% lower in the proposed method compared to the ISODATA method when the initial cluster centers are set randomly. It is also shown that the error of the proposed method is much smaller than that of the k-means method (ICCD) and the single linkage method.

B. Satellite Imagery Data used

Clustering was performed using a part (Fig. 3) near Kashima city in Japan extracted from satellite images of the Saga area taken by ASTER / VNIR: Visible and Near Infrared Radiometer on December 7, 2004. The sampling area shown in the figure was set, and 30 × 30 identical cluster image data individuals were set. Clusters of these individuals correspond to four cluster types: sea, plants, freshwater, and urban areas. The dimension of the image data individual is 3, and the maximum value of each dimension is 255. The image is shown in Fig. 4.

TABLE II. CLUSTERING ERROR OF THE FOUR CLUSTERING METHODS FOR FIVE DATASETS OF UCI REPOSITORY

Error (%)	Proposed	k-means	ISODATA	Single linkage
Iris	10.67	10.67	21.33	32.00
Wine	4.49	5.62	34.65	61.24
Ruspini	0.00	0.00	12.43	0.00
Fossil	4.60	4.60	34.42	13.79
New thyroid	13.02	13.95	31.63	29.77

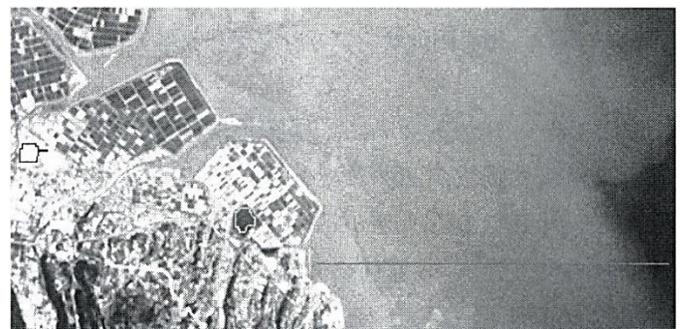


Fig. 3. Satellite Images Around Kashima City, Saga Prefecture in Japan (Band: 1,2,3 N).

⁴ <http://mllearn.ics.uci.edu/MLRepository.html>

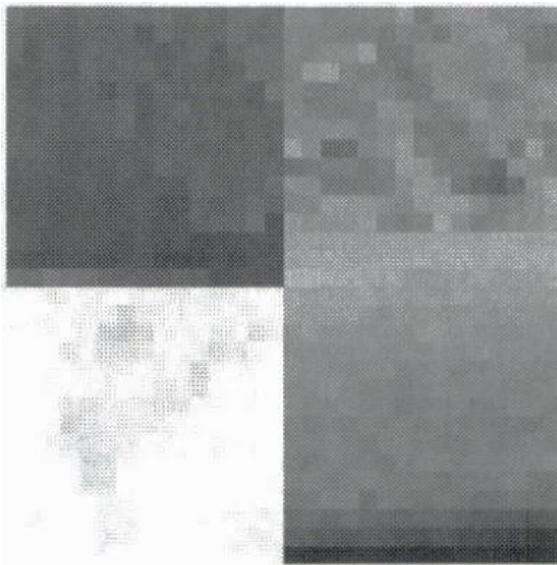


Fig. 4. Sampling Image Representing a Cluster.

The proposed method and the conventional ISODATA method are applied to the satellite image, and the cluster results are compared. In the conventional ISODATA method, the result based on the Ward method is used for the initial cluster center. The results are shown in Table III and Fig. 5. As shown in Fig. 5 of the clustered results of the proposed method and ISODATA, there are some clustering errors at the bottom right of the clustered result of ISODATA while there is no error for the proposed method.

As can be seen from this table, it is shown that the proposed method is also effective for satellite data in which the probability density function of the target data individual is a concave function. Also, from the center of the original image in Fig. 6(a), a part of $32 * 32$ pixels include the above four categories. Fig. 6(a) is extracted to perform clustering by the proposed method and the conventional ISODATA method. Applied and compared the cluster results. The results of clustering are shown in Fig. 6(b) and 6(c).

TABLE III. RESULTS OF EXPERIMENTS USING SATELLITE IMAGES

Method	Proposed	ISODATA
Error (%)	0.00	3.89
Elapsed time (s)	25567.445	1.134

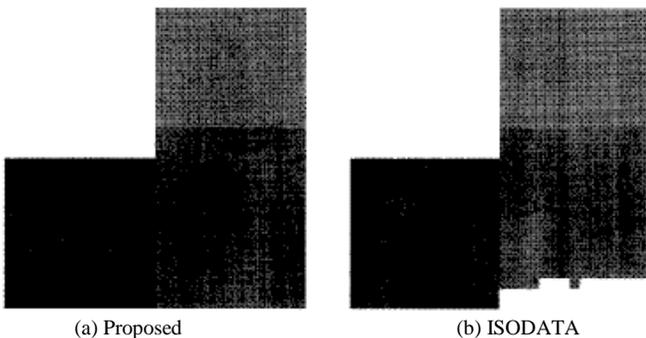


Fig. 5. Clustered Results of Proposed Method and ISODATA.



Fig. 6. Original Satellite Image and the Clustered Results of the Conventional ISODATA and the Proposed Method.

In comparison with the ISODATA method, which does not form a cluster corresponding to this distribution, it can be seen that the proposed method forms an appropriate cluster.

It can be seen that the proposed method also gives better results for satellite image data than the conventional ISODATA method, especially for classification of fresh water.

V. CONCLUSION

From the results of clustering experiments using simulation, UCI repository dataset and ASTER / VNIR images, the proposed method was superior to the conventional method in all cases. It was found that the proposed method has a higher degree of separation than the conventional method even when the probability density function of the clustering target data individual is a concave function, and a good clustering result is obtained. Therefore, the proposed method overcomes the problem that accurate clustering cannot be expected when it is difficult to set the parameters of the ISODATA method (initial cluster center and threshold of division / fusion: split / merge) and when the probability density function of the ISODATA method is concave.

Although the calculation time of the proposed method is longer than that of the conventional method by the amount of parameter setting by the real-valued genetic algorithm, the cluster accuracy is significantly improved, and it is effective when the cluster accuracy is important.

VI. FUTURE RESEARCH WORKS

Further research works are required for validation of the proposed clustering method with the other satellite imagery data. Furthermore, the alternative optimization algorithms of

GA such as sand cat swarm optimization (SCSO), Grey Wolf Optimizer (GWO) algorithm, Moth-flame optimization algorithm, etc. have to be tried for the parameter selection of the proposed clustering based on the modified ISODATA clustering.

ACKNOWLEDGMENT

The author would like to thank Prof. Dr. Hiroshi Okumura and Prof. Dr. Osamu Fukuda of Saga University for their valuable comments and suggestions.

REFERENCES

- [1] R.J. Williams, Simple statistical gradient-following algorithms for connectionist reinforcement learning, *Machine Learning*, 8, 3-4, 229-256, 1992.
- [2] C.Yi-tsuu, *Interactive Pattern Recognition*, Marcel Dekker Inc., New York and Basel, 1978.
- [3] John H.Holland, *Adaptation in Natural and Artificial Systems*, Ann Arbor, The University of Michigan Press, 1975.
- [4] John H.Holland, *Adaptation in Natural and Artificial Systems*, Massachusetts Institute of Technology, 1992.
- [5] Seyyedabbasi, A., Kiani, F. Sand Cat swarm optimization: a nature-inspired algorithm to solve global optimization problems. *Engineering with Computers* (2022).
- [6] Seyyedabbasi, A., Kiani, F. I-GWO and Ex-GWO: improved algorithms of the Grey Wolf Optimizer to solve global optimization problems. *Engineering with Computers* 37, 509-532 (2021).
- [7] Seyyedabbasi, A., Moth-flame optimization algorithm: A novel nature-inspired heuristic paradigm, *Knowledge-Based Systems*, Volume 89, November 2015, Pages 228-249, 2015.
- [8] Kohei Arai, *Basic Theory of Pattern Recognition*, Academic Book Publishing, 1999.
- [9] Ono Isao, Sato Hiroshi, Kobayashi Shigenobu, Function optimization by real-valued GA using unimodal normal distribution crossover UNDX, *Journal of Japan Society for Artificial Intelligence*, 14, 6, 1999.
- [10] Takagi Mikio, Shimoda Y., Arai Kohei, *Image Analysis Handbook*, The University of Tokyo Press, pp.641-684, 1991.
- [11] L.J. Eshleman and J.P. Scha. er, Real-Coded GA and internal schemata, *Foundations of Genetic Algorithms*, 2, 187, 1993.
- [12] E.G.M. de Lacerda, *Model Selection of RBF Networks via Genetic Algorithms*, Pernambuco Federal University Informatics Center (Report), 2003.
- [13] UCI Repository, <http://www.sgi.com/tech/mlc/db/>.
- [14] S Bandyopadhyay, An automatic shape independent clustering technique, *Machine Intelligence Unit*, January 06 *Pattern Recognition*, Vol.37, No.1, 2004.
- [15] Kohei Arai and Ali Ridho Barakba, Method for shape independent clustering in case of numerical clustering together with condensed clustering, *Proc. of the SCI symposium 2004*.
- [16] Akira Yoshizawa and Kohei Arai, Influence due to geomorphology on context genetic Algorithm Clustering, *Transaction of the Japanese Geomorphological Union*, Vol.22, No.4, 23-28, 2001.
- [17] Kohei Arai, Learning processes of image clustering method with density maps derived from Self-Organizing Mapping(SOM), *Journal of Japanese Society of Photogrammetry and Remote Sensing*, 43, 5, 62-67, 2004.
- [18] Kohei Arai, Non-linear merge and split method for image clustering, *Journal of Japanese Society of Photogrammetry and Remote Sensing*, 43, 5, 68-73, 2004.
- [19] Ali Ridho Barakbah and Kohei Arai, Revised pattern of moving variance for acceleration of automatic clustering, *Electric and Electronics Polytechnics in Surabaya EEPIS Journal*, 9, 7, 15-21 (2004).
- [20] Kohei Arai, I.N.Abdullah, H.Okumura, Automatic detection method for clustered micro calcification in mammogram image based on statistical texture features, *International Journal of Advanced Computer Science and Applications*, 3, 5, 7-11, 2012.
- [21] Kohei Arai, X.Q. Bu, Comparative study between the proposed GA based ISODATA clustering and the conventional clustering methods, *International Journal of Advanced Computer Science and Applications*, 3, 7, 125-131, 2012.
- [22] Kohei Arai, Image clustering method based on density maps derived from Self Organizing Mapping: SOM, *International Journal of Advanced Computer Science and Applications*, 3, 7, 102-107, 2012.
- [23] Kohei Arai, Clustering method based on Messy Genetic Algorithm: GA for remote sensing satellite image clustering, *International Journal of Advanced Research in Artificial Intelligence*, 1, 8, 6-11, 2012.
- [24] Kohei Arai, Visualization of learning process for back propagation Neural Network clustering, *International Journal of Advanced Computer Science and Applications*, 4, 2, 234-238, 2013.
- [25] Kohei Arai, Indra Nugraha Abdullah, Hiroshi Okumura, Rie Kawakami, Improvement of automated detection method for clustered micro calcification base on wavelet transformation and support vector machine, *International Journal of Advanced Research in Artificial Intelligence*, 2, 4, 23-28, 2013.
- [26] Kohei Arai, Image clustering method based on Self-Organization Mapping: SOM derived density maps and its application for Landsat Thematic Mapper image clustering, *International Journal of Advanced Research in Artificial Intelligence*, 2, 5, 22-31, 2013.
- [27] Kohei Arai, Cahya Rahmad, Comparative study between the proposed shape independent clustering method and the conventional method (k-means and the others), *International Journal of Advanced Research in Artificial Intelligence*, 2, 7, 1-5, 2013.
- [28] Kohei Arai, Genetic Algorithm utilizing image clustering with merge and split processes which allows minimizing Fisher distance between clusters, *International Journal of Advanced Research in Artificial Intelligence*, 2, 9, 7-13, 2013.
- [29] Kohei Arai, Fisher distance based GA clustering taking into account overlapped space among probability density functions of clusters in feature space, *International Journal of Advanced Research in Artificial Intelligence*, 2, 11, 32-37, 2013.
- [30] Ali Ridho Barakbah, Kohei Arai, Centronit: Initial Centroid Designation Algorithm for K-Means Clustering, *EMITTER: International Journal of Engineering Technology*, 2, 1, 50-62, 2014.
- [31] Kohei Arai, Pursuit Reinforcement Competitive Learning: PRCL Based Online Clustering with Tracking Algorithm and Its Application to Image Retrieval, *International Journal of Advanced Research on Artificial Intelligence*, 5, 9, 9-16, 2016.

AUTHORS' PROFILE

Kohei Arai, He received BS, MS and PhD degrees in 1972, 1974 and 1982, respectively. He was with The Institute for Industrial Science and Technology of the University of Tokyo from April 1974 to December 1978 also was with National Space Development Agency of Japan from January, 1979 to March, 1990. During from 1985 to 1987, he was with Canada Centre for Remote Sensing as a Post Doctoral Fellow of National Science and Engineering Research Council of Canada. He moved to Saga University as a Professor in Department of Information Science on April 1990. He was a councilor for the Aeronautics and Space related to the Technology Committee of the Ministry of Science and Technology during from 1998 to 2000. He was a councilor of Saga University for 2002 and 2003. He also was an executive councilor for the Remote Sensing Society of Japan for 2003 to 2005. He is a Science Council of Japan Special Member since 2012. He is an Adjunct Professor of University of Arizona, USA since 1998. He also is Vice Chairman of the Science Commission "A" of ICSU/COSPAR since 2008 then he is now award committee member of ICSU/COSPAR. He wrote 55 books and published 680 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 Pre-trained Neural Network to Predict Alzheimer's Disease at an Early Stage

Ragavamsi Davuluri, Ragupathy Rengaswamy

Department of Computer Science and Engineering
Annamalai University, Chidambaram, Tamil Nadu - 608001

Abstract—Alzheimer's disease (AD), which is a neuro associated disease, has become a common for past few years. In this competitive world, individual has to perform lot of multi tasking to prove their efficiency, in this process the neurons in the brain gets affected after a while i.e., "Alzheimer's Disease". Existing models to identify the disease at early stage has taken the individuals speech as input then they are converted into textual transcripts. These transcripts are analyzed using neural network approached by integrating them with NLP techniques. These techniques failed in designing the model which can process the long conversation text at faster rate and few models are unable to recognize the replacement of the unknown words during the translation process. The proposed system addresses these issues by converting the speech obtained into image format and then the output "Mel-spectrum" is passed as input to pre-trained VGG-16. This process has greatly reduced the pre-processing step and improved the efficiency of the system with less kernel size architecture. The speech to image translation mechanism has improved accuracy when compared to speech to text translators.

Keywords—Mel-spectrum; VGG-16; ADAM optimizer; softmax; flatten layers; ReLU

I. INTRODUCTION

Alzheimer's disease is a progressive brain disease that gradually deteriorates memory and thinking abilities, as well as the ability to do even the most fundamental tasks. The majority of people with late-onset type symptoms are in their mid-60s when they get the disease [6]. Early-onset Alzheimer's disease is extremely rare and occurs between the ages of 30 and 60. The most prevalent cause of dementia in elderly people is Alzheimer's disease [20]. Memory issues are usually one of the early signs of Alzheimer's disease, though the severity of the symptoms varies from person to person [7]. Other areas of thinking, such as finding the proper words, vision/spatial difficulties, and impaired reasoning or judgments, may also indicate Alzheimer's disease in its early stages. There is no cure or treatment for Alzheimer's disease that affects the disease process in the brain. Complications from severe loss of brain function, such as dehydration, malnutrition, or infection, result in death in advanced stages of the disease [13]. Alzheimer's disease can be detected using a machine learning approach, which involves the use of various machine learning algorithms [18]. Furthermore, the patient's severity level will be predicted in percentages, and the percentage levels will be divided into several categories. The importance of early detection in Alzheimer's disease management cannot be overstated [14]. Convolutional neural network (CNN) is one of the deep-

learning algorithms that have been used to detect structural brain alterations on magnetic resonance imaging (MRI) because of its high efficiency in automated feature learning [15]. Many additional deep learning methods are being utilized to diagnose Alzheimer's disease [17] and even pre-trained models can be used in the detection of Alzheimer's disease [19].

The proposed model uses the audio dataset to improve the accuracy of the model by analyzing the live streaming data in case of Alzheimer's disease prediction [8]. The model produces a spectrogram from audio files, which produces visualization of signal strength in the form of 2D graph. The model needs to do two pre-processing steps before converting it into image [23].

The audio data is stored in the digitized format. Any machine learning algorithm is difficult to work with this digital form. So machine needs sampling mechanism to convert the digital data into analog data [9]. Sampling technique transforms the signal with respect to time into numerical values by identifying the difference between two consecutive samples of audio segments [10].

The obtained sample may contain noisy values i.e., at few intervals the obtained time signals may have amplitude has zero, which represents the state of silence [21]. The model address this issue by performing quantization that helps them to replace silence with nearest precision value and then it normalizes the data to have values in between -1 to 1 [24].

The model needs frequencies to identify the voice modulation, so it applies Fourier transformations to convert the time signals into individual frequencies. The frequencies can be produced into two ways namely FFT & STFT. The model employs STFT because it can efficiently convert the 1-dimensional data into 2-dimensional data where horizontal axis represents time and vertical axis represents frequency [11]. In short note, the sound waves are segmented into smaller chunks and few of them might overlap. In the final stage, each frequency amplitude in decibels is stored as a "pixel" value of the image. This conversion process is known as "Mel Scale" [12]. These pixels need to be stored in the form of vectors. The spectrogram represented in Mel Scale is a Mel Spectrogram [22].

II. LITERATURE SURVEY

Rohanian et al. proposed Multi-modal fusion on sequential modeling using LSTM technique. These fusion models are good in handling the lexical information. Since this context

changes with time, the model uses gating concept. This receives inputs from two different models i.e., (audio and text) and combines them into single unit by eliminating the noises from the audio data acquired. All the higher layers utilize non-linear functions but the lower layers use linear functions. The non-linear is transformed into linear by attaching the carriers integrated with self repairing system. The model hyper tunes the LSTM to identify the error co-efficient at early stage [1].

Raghavendra et al. designed embedded techniques related to speech and fine tune those using BERT models. The model extracts the essential features by constructing the x-vectors, which is a deep neural vector. In the next step, Mel-frequency coefficient is computed by embedding all the global pooling layers. The encoder of the ResNet stores the signals in the encrypted format but at decoder side it implements a pre-trained VoxCeleb1. The performance of this model is estimated using the MMSE because most of the characteristics are relevant to prosody features. These features are hard to maintain, so BERT model using its self attention layer which converts the linguistics elements into embeddings [2].

Ning Wang et al. implemented Attention Network by collecting real time data from Google Speech Recognizer API. The model extracts four different features using four different networks. The model for extracting the frame level features of the linguistic, it implements VGG network because it is good in handling the embossed features. FREQ commands convert the received audio into transcript form. NLTK takes care of the converted text but the context of each sentence is extracted using the Universal Sentence Embedding technique [3]. The model implements dilated CNN layers instead of 2D-CNN layers because multi head component associated with the embedded layers can represent the encoded representation very efficiently [16].

Amish et al. designed a framework BERT integrated CNN to extract the embedding text associated with the context. The audio dataset is initially segmented into shorter clips and then Mel-spectrum is generated for those segments. The model utilized Fast Text-CNN to generate transcripts and clusters are formed based on the common word vectors. In this model, instead of encoding all the sentences, it encrypts only Out-of-Vocabulary words using sentence BERT technique. During the text processing phase, the probability of segment is compared against the probability of entire transcript and then combined to form a fusion. Finally, a concatenation embedding model is designed to treat the each text segments separately [4].

Zhaoci Liu et al. worked with bottle neck features generated from augment images. The model doesn't convert the audio signals into transcripts because the designed extractor creates temporary intermediate form of representation. The audio signals are divided into set of window frames, from which both local and global context information is extracted. This information is stored as a sequence in LSTM and attention pooling is applied to perform the classification. The validation of the model is performed using query system, which is designed as feed forward network with sigmoid activation function and the input is obtained in the form of key-value pair [5].

III. PROPOSED METHODOLOGY

Most of the researchers analyzed the Alzheimer's disease either from images or from the .csv file extracted from the images. Few researchers, which are stated in Table I, worked on predicting system using audio dataset, which is converted into textual format. The accessing and processing of text using NLP techniques takes a lot of time, which makes the system to take late decisions. This issue is resolved in this paper by transforming the speech into images and processing them using pre-trained model. The entire process is illustrated in Fig. 1.

TABLE I. IDENTIFICATION OF LIMITATIONS FROM EXISTING MODEL

S.No	Author	Algorithm	Merits	Demerits	Accuracy
1	Rohanian et al. [1]	Multi-modal fusion	The feed-forward helps the model to transfer the data quickly	The model can be extended by introducing the bio-markers	79.20
2	Raghavendra et al. [2]	BERT	Embedding BERT and X-vector has solved the modulation frequencies very effectively during speech recognition	Adaptable BERT integrated with LM interpolation will refine the predictions associated with target values	84.51
3	Ning Wang et al. [3]	Attention Network	Instead of single feature extraction, the model extracted multiple features from different sources. So, this model achieves less misclassification rate	The model doesn't apply any segmentation technique to process the speech with respect to time	80.28
4	Amish et al. [4]	Multi Modal Sequence	The FastText CNN has the capability to generate transcripts for unknown words also	The model has implemented late fusion technique to combine the text segments which consumes more memory to activate the cells	85.30
5	Zhaoci Liu et al. [5]	Masked data augmentation	Since the bottleneck features extracts the low level values, it requires less space and time to encrypt	Usage of query system for validating the model made the process complicated	82.59

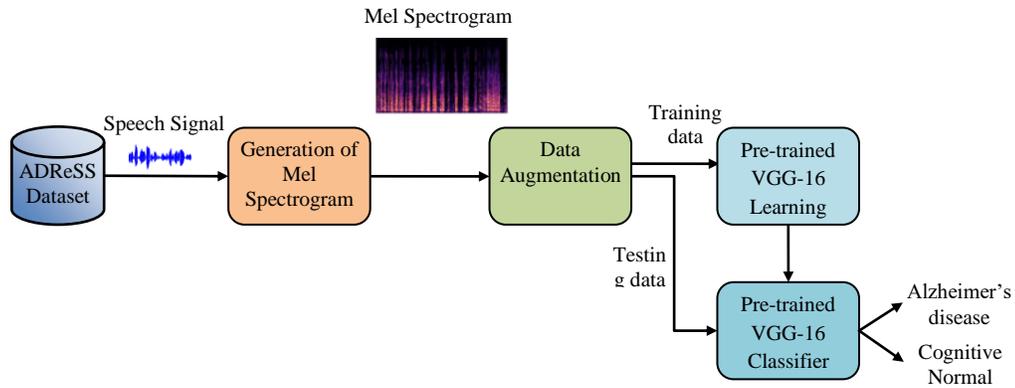


Fig. 1. Overall Architecture of Proposed System.

A. Conversion of Speech Signals to Mel-spectrum Images

The proposed research focused to analyze the disease from the speech, which are basically classified into Alzheimer’s disease (AD) and Cognitive normal (CN). Initially the model has constructed the Mel spectrogram from the speech files as represented in Fig. 2.

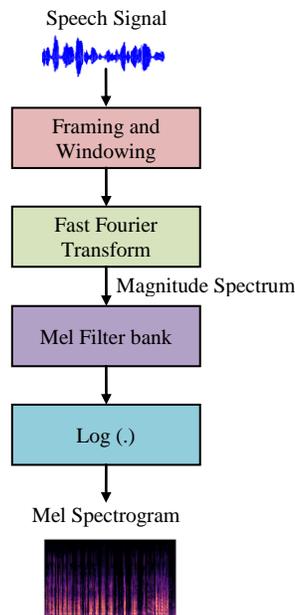


Fig. 2. Generation of Spectrogram Images from Audio Dataset.

The approach employs samples of air pressure over time to digitally represent a speech signal. The voice signal is transferred from the time domain to the frequency domain using the fast Fourier transform, and the system uses overlapping windowed portions of the speech stream. The technology converts the y-axis (frequency) to a log scale and the color dimension (amplitude) to decibels to create the

spectrogram. The y-axis (frequency) was mapped onto the Mel scale to generate the Mel spectrogram. The model applies a traditional Short Time Fourier Transformation (STFT); this is a very flexible class to represent time and frequency distribution from the processed speech signals. The identification of pitch variation is the major hyper tuning point. The computation is presented in equation (1).

$$STFT_m(input) = \sum_{n=-\infty}^{\infty} input_n * [H(n) - R_m] * e^{-j} \quad (1)$$

where, $input_n$ denotes input signal recognized at time ‘n’ $H(n)$ denotes N-length Hamming function applied on sliding window

R_m Represents hop size in between the m-size sliding window e^{-j} is a threshold multiplication function

Humans do not perceive frequencies on a linear scale, according to research. Lower frequency differences are easier to notice than higher frequency variances. Humans can readily distinguish between 500 and 1000 Hz, but we will struggle to distinguish between 10,000 and 10,500 Hz, despite the fact that the distance between the two pairs is same. As a result, we’ll utilize the Mel scale, which is a logarithmic scale based on the idea that equal lengths on the scale correspond to the same perceptual distance. Conversion from frequency (f) to Mel scale (m) is given in equation (2).

$$m = 2595 \cdot \log\left(1 + \frac{f}{500}\right) \quad (2)$$

A Mel Spectrogram is a spectrogram that converts frequencies to Mel scale. For the creation of log-Mel spectrograms, we chose 40 Mel filter banks. The availability of 40 Mel filter banks allows us to use pre-trained models such as VGG16 in the future. A Hamming window with a size of 2048 samples is used. With a hop-length of 1024 samples (the amount of samples between subsequent frames) and a sampling rate of 44.1 kHz. In the FFT calculation, the number of points is also 2048. The log-Mel spectrograms are separated into overlapping chunks of 100 frames each encompassing a length of 23ms to achieve this.

Algorithm for Construction of Mel Spectrum

```
Input: Audio Dataset with binary class labels, AD_Binary
Output: Mel-spectrum Generation
Begin
1. Set the path of the output folder to store the images of
spectrograms
2. ad_binary_labels←[“AD”, “CN”]
3. for i in ad_binary_labels: for j in path:
    a. audio_time_series,audio_sr←load(j,sr=None)
    // load the speech data in time series format with the
    specified sampling rate
    b. speech_frequency←stft(speech_time_series)
    c. speech_magintude, speech_phase ← magphase
    (speech_frequency)
    d. mel_scale_speech ← melspectrogram
    (S=speech_magintude,sr=speech_sr)
    e. mel_speech← amplitude_to_db
    (mel_scale_speech,ref=“minimum”)
    f. specshow(mel_speech,sr=speech_sr)
End
```

B. Disease Classification using VGG-16

The proposed system applies VGG-16 pre-trained model on generated spectrogram images to identify the disease. The model has chosen VGG network because of its simplicity nature even though it has huge parameter to hyper tune. The basic thumb rule for any deep learning algorithms is “More the training data more the accuracy”, but the training dataset has got fewer images from the speech signals. This issue is resolved by the model initially by applying basic image manipulation operations to increase the size of the generated spectrogram images as shown in Fig. 3.

The model has customized the few operations in the generator module to minimize the error rate. The basic customized operation is “pre-processing” unit because it denoises the images, which is a basic challenge faced by any of the computer vision applications. In this model, images are generated from the speech signals; therefore there are high chances to get noisy images due to sudden voice modulation from the external factors. The remaining customization operations are elaborated in Table II.

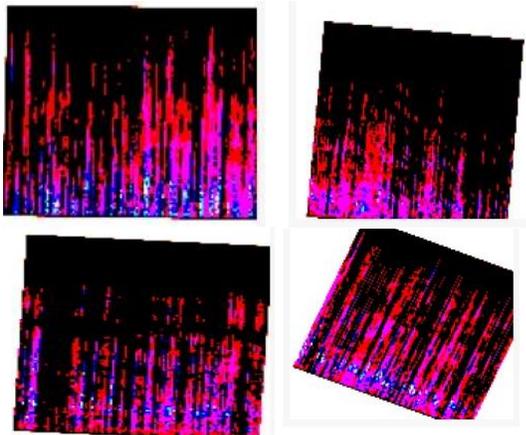


Fig. 3. Synthetic Data Created using the Image Data Generator Module.

The main advantages of the generator module lies in passing the multiple operations performed on the image are sent directly to the neural network instead of storing them in a temporary buffer or memory. Finally, the images are rescaled to 150 and divided them into 32 batch size. These synthesized images are passed as input to the VGG-16 neural network, whose task is achieved in two stages. In the first stage, it identifies the objects from the generated signals from the pre-defined classes available. In the second stage, the model contains 1000 class labels, so it has to classify the generated image from the customized class label. The model implemented its validation across Image Net dataset. The model implements only a kernel filter with size 3×3. The overall layers and their configurations for the VGG-16 are presented in Table III.

The entire architecture of the model is divided into 5 blocks of Ensemble 3-Dimensional CNN layers and 1 block of Fully Connected (or) Dense Layer. It takes standard input of size 224×224 with 3 dimensions. The first two blocks contains 2 layers of CNN with max pooling layer then remaining blocks contains 3 layers of CNN with max pooling. The final block contains two fully connected layers and one softmax layer.

TABLE II. DESCRIPTION OF IMAGE MANIPULATION CUSTOMIZATION OPERATIONS IN GENERATOR MODULE

S.No	Parameter Name	Description	Initialized Value
1	preprocessing_function	It removes the noise from generated images	Processed images from VGG16
2	Rotation_range	Some random images are rotated to 40 degrees angle	40
3	Width_shift	The images are translated 20% horizontally based on total width	0.2
4	Height_shift	The images are translated 20% vertically based on total height	0.2
5	Shear_range	It transforms the point to a particular	0.2
6	Zoom_range	The inside image contents are zoomed to 20%	0.2
7	Horizontal_flip	It randomly flips the half images of the dataset	True
8	Fill_mode	It fills the newly created pixels with nearest pixel values	Nearest

TABLE III. VGG-16 CONFIGURATION

S.No	Layer Name	Dimensions	Activation Function	Count
1	Convolution Layer	Initially it starts with 64 then it enhances up to 512	ReLU	13
2	Max Pooling Layer	From 75 × 75 × 64 to 4 × 4 × 512	-	5
3	Dense Layer	1×1×4096 is converted into 1×1×1000	ReLU	2

The customization of the pre-trained model is represented in the below section:

- 1) The model acquired the input images in the 2-Dimensional but VGG-16 requires in three dimensional. So it changes the input shape from 2-D to 3-D by passing additional parameters.
- 2) The model adjusted the weights of the neural network based on the “ImageNet” dataset.

3) The augment images have the shape of 150×150 but the VGG-16 accepts 224×224 . So, they include top attribute has assigned with false value.

4) In general VGG-16 requires softmax layer for multi classification but the model dataset has binary class labels. So, the model implements the softmax layer by flattening the layer. The overall architecture of VGG-16 is presented in Fig. 4.

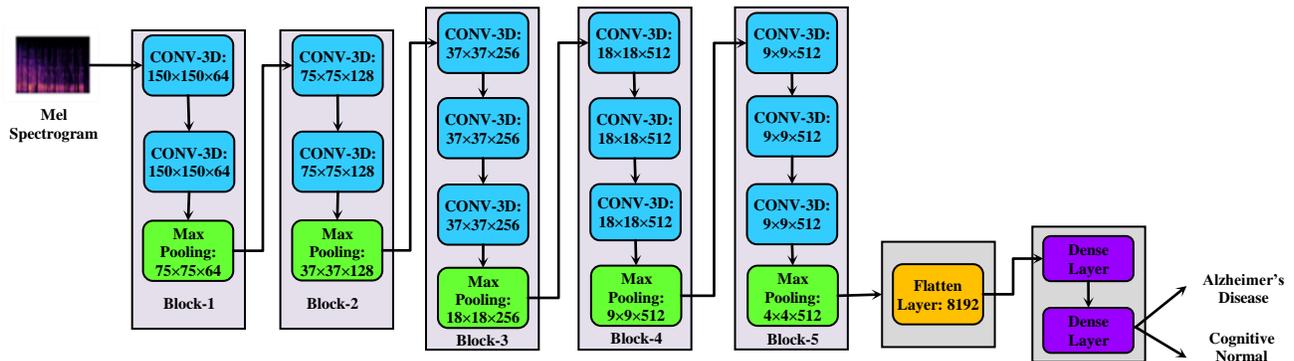


Fig. 4. Pre-trained Architecture of VGG-16.

IV. RESULT AND DISCUSSION

This model considers dataset from the publicly available repositories and uses Google Laboratory as the execution environment because the model needs GPU's to work with speech signals. The speech signals are labeled as “AD” and “CN”. Alzheimer's is a group of neurodegenerative disorders characterized by a steady and long-term decline in cognitive function. Because age is the most important risk factor for AD, it affects the elderly the most. Because of the global severity of the situation, institutions and researchers are putting significant resources towards Alzheimer prevention and early detection, with an emphasis on disease progression. Cost-effective and scalable approaches for detecting Alzheimer disease in its most mild manifestations are needed. While several studies have investigated speech and language features for Alzheimer's disease and proposed various signal processing and machine learning methods for this task, the field still lacks balanced and standardized data sets on which these different approaches can be systematically compared. The ADReSS challenge dataset includes CN and AD patients' speech recordings, transcripts, and metadata (age, gender, and MMSE score). The dataset is balanced in terms of age, gender, and the number of CN vs. AD patients, with 78 patients in each class. The speech data is converted into Mel spectrograms. The outputs for the individual modules are represented in the below section.

Fig. 5(a) and 5(b) represents the generation of spectrogram from the speech signals for both the class labels. The graphs are represented by using the voice frequency obtained from the speech signals.

Fig. 6 represents the trainable and non-trainable parameters of VGG-16 by customizing the necessary parameters and

layers. In general, neural networks need more parameters to train the model but due to usage of pre-trained models, the trainable parameters got reduced.

Fig. 7 represents the training phase of the model in pre-defined Epochs values. With the increase of Epoch, there might be increase or decrease of the accuracy. So, the model saves the highest accuracy as the best model. It updates the checkpoints if and only if the model gets better accuracy than the previous value. The model wants to prove the state of accuracy by comparing the standard CNN with VGG-16 pre-trained model. So, it represented the initial output training results of CNN in Fig. 8.

Fig. 9 represents the accuracy values obtained by VGG-16 at different Epochs levels. From Fig. 10, the model can clearly state that the VGG-16 has performed very well than CNN by comparing from the initial Epochs.

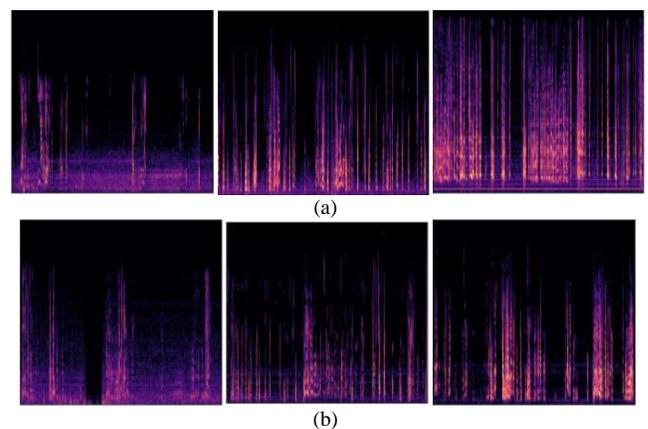


Fig. 5. Mel Spectrum. (a) AD Class Label. (b) CN Class Label.

Layer (type)	Output Shape	Param #
input_1 (InputLayer)	[(None, 150, 150, 3)]	0
block1_conv1 (Conv2D)	(None, 150, 150, 64)	1792
block1_conv2 (Conv2D)	(None, 150, 150, 64)	36928
block1_pool (MaxPooling2D)	(None, 75, 75, 64)	0
block2_conv1 (Conv2D)	(None, 75, 75, 128)	73856
block2_conv2 (Conv2D)	(None, 75, 75, 128)	147584
block2_pool (MaxPooling2D)	(None, 37, 37, 128)	0
block3_conv1 (Conv2D)	(None, 37, 37, 256)	295168
block3_conv2 (Conv2D)	(None, 37, 37, 256)	590080
block3_conv3 (Conv2D)	(None, 37, 37, 256)	590080
block3_pool (MaxPooling2D)	(None, 18, 18, 256)	0
block4_conv1 (Conv2D)	(None, 18, 18, 512)	1180160
block4_conv2 (Conv2D)	(None, 18, 18, 512)	2359808
block4_conv3 (Conv2D)	(None, 18, 18, 512)	2359808
block4_pool (MaxPooling2D)	(None, 9, 9, 512)	0
block5_conv1 (Conv2D)	(None, 9, 9, 512)	2359808
block5_conv2 (Conv2D)	(None, 9, 9, 512)	2359808
block5_conv3 (Conv2D)	(None, 9, 9, 512)	2359808
block5_pool (MaxPooling2D)	(None, 4, 4, 512)	0
flatten (Flatten)	(None, 8192)	0
dense (Dense)	(None, 2)	16386

=====
 Total params: 14,731,074
 Trainable params: 16,386
 Non-trainable params: 14,714,688
 =====

Fig. 6. Summary of VGG-16.

2/2 - 11s - loss: 0.3769 - accuracy: 0.9375 - 11s/epoch - 5s/step
 Epoch 24/30
 WARNING:tensorflow:Can save best model only with val_loss available, skipping.
 2/2 - 11s - loss: 0.6510 - accuracy: 0.9375 - 11s/epoch - 5s/step
 Epoch 25/30
 WARNING:tensorflow:Can save best model only with val_loss available, skipping.
 2/2 - 11s - loss: 0.3480 - accuracy: 0.9583 - 11s/epoch - 5s/step
 Epoch 26/30
 WARNING:tensorflow:Can save best model only with val_loss available, skipping.
 2/2 - 11s - loss: 0.5336 - accuracy: 0.8958 - 11s/epoch - 5s/step
 Epoch 27/30
 WARNING:tensorflow:Can save best model only with val_loss available, skipping.
 2/2 - 11s - loss: 0.9995 - accuracy: 0.8542 - 11s/epoch - 5s/step
 Epoch 28/30
 WARNING:tensorflow:Can save best model only with val_loss available, skipping.
 2/2 - 11s - loss: 0.3081 - accuracy: 0.9583 - 11s/epoch - 5s/step
 Epoch 29/30
 WARNING:tensorflow:Can save best model only with val_loss available, skipping.
 2/2 - 11s - loss: 1.0050 - accuracy: 0.9375 - 11s/epoch - 5s/step
 Epoch 30/30
 WARNING:tensorflow:Can save best model only with val_loss available, skipping.
 2/2 - 11s - loss: 0.6041 - accuracy: 0.9167 - 11s/epoch - 5s/step
 Training completed in time: 0:08:30.489462

Fig. 7. Few Training Iterations of the Proposed Model.

Epoch 1/10
 3/3 [=====] - 22s 10s/step - loss: 0.5361 - accuracy: 0.7945
 Epoch 2/10
 3/3 [=====] - 2s 402ms/step - loss: 0.4808 - accuracy: 0.8356
 Epoch 3/10
 3/3 [=====] - 2s 413ms/step - loss: 0.4402 - accuracy: 0.8356
 Epoch 4/10
 3/3 [=====] - 2s 678ms/step - loss: 0.4568 - accuracy: 0.8356
 Epoch 5/10
 3/3 [=====] - 2s 680ms/step - loss: 0.4255 - accuracy: 0.8356
 Epoch 6/10
 3/3 [=====] - 2s 412ms/step - loss: 0.4219 - accuracy: 0.8356
 Epoch 7/10
 3/3 [=====] - 2s 645ms/step - loss: 0.4092 - accuracy: 0.8356
 Epoch 8/10
 3/3 [=====] - 2s 409ms/step - loss: 0.3927 - accuracy: 0.8356
 Epoch 9/10
 3/3 [=====] - 2s 689ms/step - loss: 0.3933 - accuracy: 0.8356
 Epoch 10/10
 3/3 [=====] - 2s 410ms/step - loss: 0.3557 - accuracy: 0.8356

Fig. 8. Training Process using Standard CNN.

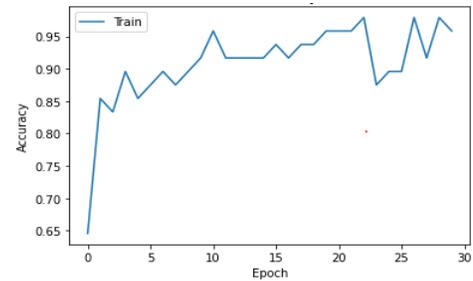


Fig. 9. Accuracy Graph Obtained by the VGG-16.

With reference to Table I and from the above figures, the Fig. 10 plots the accuracies obtained by different existing models along with proposed and standard CNN architecture to project the performance of the model. X-axis represents the approaches and Y-axis represents the accuracy obtained.

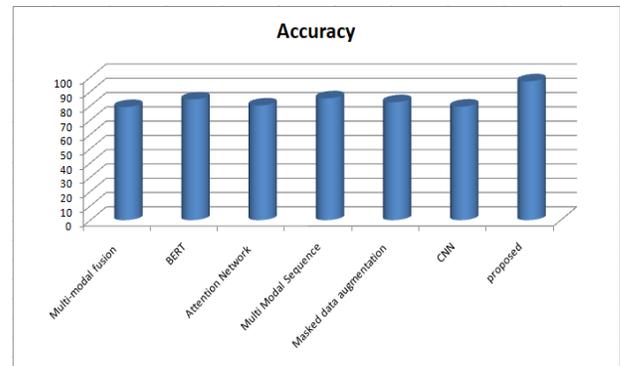


Fig. 10. Comparison Analysis.

V. CONCLUSION

In this paper, the model has recognized Alzheimer's disease at early stage effectively by creating the synthetic dataset of Mel spectrum images then these are acted as input for the ImageNet trained VGG-16 system. The model has opted pre-trained model instead of CNN because the dataset contains different modulation signals with noise. The design becomes complicated because to extract essential features from the

different modulations, more number of layers is required. It requires an efficient back propagation system to update the weights accurately. The numbers of trainable parameters are 8,53,145 in CNN where as the number of trainable parameters are 16,386 in proposed network. The misclassification rate and accuracy are also got affected because of the transfer learning process. As a conclusion remarks, it can be stated that variations in CNN model work well on the images because of its implicit feature extraction process.

REFERENCES

- [1] M. Rohanian, J. Hough, and M. Purver, "Multi-Modal Fusion with Gating Using Audio, Lexical and Disfluency Features for Alzheimer's Dementia Recognition from Spontaneous Speech," *Interspeech*, pp. 2187–2191, October 2020.
- [2] P. Raghavendra, J. Cho, S. Joshi, L. Moro-Velazquez, P. Żelasko, J. Villalba, and N. Dehak, "Automatic detection and assessment of Alzheimer Disease using speech and language technologies in low-resource scenarios," *Interspeech*, pp. 3825–3829, June 2021.
- [3] N. Wang, Y. Cao, S. Hao, Z. Shao, and K. P. Subbalakshmi, "Modular Multi-Modal Attention Network for Alzheimer's Disease Detection Using Patient Audio and Language Data," *Interspeech*, August 2021.
- [4] A. Mittal, S. Sahoo, A. Datar, J. Kadiwala, H. Shalu, and J. Mathew, "Multi-Modal Detection of Alzheimer's Disease from Speech and Text," *arXiv*, July 2021.
- [5] Z. Liu, Z. Guo, Z. Ling, and Y. Li, "Detecting Alzheimer's Disease from Speech Using Neural Networks with Bottleneck Features and Data Augmentation," *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 7323–7327, June 2021.
- [6] D. Ragavamsi, and R. Ragupathy, "A Survey of Different Machine Learning Models for Alzheimer Disease Prediction," *International Journal of Emerging Trends in Engineering Research*, vol.8, pp. 3328–3337, July 2020.
- [7] D. Ragavamsi, and R. Ragupathy, "Identification of Alzheimer's Disease Using Various Deep Learning Techniques—A Review," *Smart Innovation Systems and Technologies*, vol. 265, pp. 485–498, December 2021.
- [8] I. Vigo, L. Coelho, and S. Reis, "Speech- and Language-Based Classification of Alzheimer's Disease: A Systematic Review," *Bioengineering*, vol. 9, 2022.
- [9] Y. Qin, W. Liu, Z. Peng, S. Ng, J. Li, H. Hu, and T. Lee, "Exploiting Pre-Trained ASR Models for Alzheimer's Disease Recognition Through Spontaneous Speech," *arXiv*, vol. 9, January 2022.
- [10] J. Laguarda, and B. Subirana, "Longitudinal Speech Biomarkers for Automated Alzheimer's Detection," *Frontiers in Computer Science*, vol. 3, April 2021.
- [11] S. Al-Shoukry, T. H. Rassem, and N. M. Makbol, "Alzheimer's Diseases Detection by Using Deep Learning Algorithms: A Mini-Review," *IEEE Access*, vol. 8, pp. 77131–77141, April 2020.
- [12] Q. Zhou, J. Shan, W. Ding, C. Wang, S. Yuan, F. Sun, H. Li, and B. Fang, "Cough Recognition Based on Mel-Spectrogram and Convolutional Neural Network," *Frontiers in robotics and AI*, vol. 8, May 2021.
- [13] T. Jo, K. Nho, and A. J. Saykin, "Deep Learning in Alzheimer's Disease: Diagnostic Classification and Prognostic Prediction Using Neuroimaging Data," *Frontiers in aging neuroscience*, vol. 11, August 2019.
- [14] S. Kaur, S. Gupta, S. Singh, and I. Gupta, "Detection of Alzheimer's Disease Using Deep Convolutional Neural Network," *International Journal of Image and Graphics*, January 2021.
- [15] Y. Wang, X. Liu, and C. Yu, "Assisted Diagnosis of Alzheimer's Disease Based on Deep Learning and Multimodal Feature Fusion," *Complexity*, April 2021.
- [16] J. Islam, and Y. Zhang, "Brain MRI analysis for Alzheimer's disease diagnosis using an ensemble system of deep convolutional neural networks," *Brain Informatics*, May 2018.
- [17] R. R. Janghel, "Deep-Learning-Based Classification and Diagnosis of Alzheimer's Disease," *Deep Learning and Neural Networks*, pp. 1358–1382, 2020.
- [18] D. Ragavamsi, and R. Ragupathy, "Neuro-imaging-based Diagnosing System for Alzheimer's Disease Using Machine Learning Algorithms," *Innovations in Computer Science and Engineering*, vol. 385, pp. 501–509, March 2022.
- [19] Y. Qin, W. Liu, Z. Peng, S. Ng, J. Li, H. Hu, and T. Lee, "Exploiting Pre-Trained ASR Models for Alzheimer's Disease Recognition Through Spontaneous Speech," *arXiv*, October 2021.
- [20] L. Ilias, D. Askounis, and J. Psarras, "Detecting Dementia from Speech and Transcripts using Transformers," *arXiv*, October 2021.
- [21] A Mallikarjuna Reddy, Vakulabharanam Venkata Krishna, Lingamgunta Sumalatha and Avuku Obulesh, "Age Classification Using Motif and Statistical Features Derived On Gradient Facial Images", *Recent Advances in Computer Science and Communications*, vol.13,2020.
- [22] A. Meghanani, C. S. Anoop, and A. G. Ramakrishnan, "An Exploration of Log-Mel Spectrogram and MFCC Features for Alzheimer's Dementia Recognition from Spontaneous Speech," *IEEE*, January 2021.
- [23] Y. Zhu, X. Liang, J. Batsis, and R. M. Roth, "Exploring Deep Transfer Learning Techniques for Alzheimer's Dementia Detection," *Frontiers in Computer Science*, Vol. 3, May 2021.
- [24] M. S. Syed, Z. S. Syed, E. Pirogova, and M. Lech, "Static vs. Dynamic Modelling of Acoustic Speech Features for Detection of Dementia," *International Journal of Advanced Computer Science and Applications*, vol. 11, October 2020.

A k -interpolation Model Clustering Algorithm based on Kriging Method

Guoyan Chen*, Yaping Qian
School of Mechanical Engineering
Jiangsu University of Technology, Changzhou, China

Abstract—In this work, a k -interpolation model clustering algorithm is proposed based on Kriging method, aim to partition data according to the relationship between the response of interest and input variables. Kriging method is used to describe the relationship between the response of interest and input variables. For each datum, the estimation errors of the interpolation models of the clusters are used to decide its assignment. An optimization strategy is proposed to obtain the final clustering results. The key factors of the proposed algorithm on its performance are studied through the synthetic and real-world datasets. The results show that the proposed algorithm is able to cluster the data according to the response of interest and input variables, and provides competitive clustering performance compared with the other clustering algorithms.

Keywords—Data clustering; Kriging method; k -means algorithm; interpolation model

I. INTRODUCTION

In recent years, massive data have been generated and recorded from real-world systems. The information mined from these data represents the characteristics of the real-world system, which can be used to analyze and improve the performance of the system. In most data mining tasks, it is necessary to build the performance prediction model first, aiming to accurately estimate the response of interest according to the input variables. However, the relationship between the response and input often changes greatly, which is difficult to evaluate through a unified prediction model [1-3]. Obviously, this issue can be solved by partitioning the data so that the data in the same part have a more similar relationship between response and input than the data from the other parts, and this work can be accomplished by data clustering.

Data clustering is a class of algorithms and techniques aiming to partition a dataset such that the data characteristics in the same cluster are more similar than the other clusters [4]. Many clustering algorithms have been proposed in the past decades, such as k -means algorithm, fuzzy c -means algorithm, Gaussian mixture model, and so on [5-6]. Since the k -means algorithm is easy to understand and implement, it has been widely used in many data mining tasks such as image recognition, modal analysis, and outlier detection. Shubair, and Al-Nassiri used the least square method to estimate the centers of clusters in k -means algorithm and applied the clustering algorithm in the preparation process of data streams [7]. Aldino et al. used k -means algorithm to group the corn-producing regions based on the collected data of corn crops to assist in the formulation of corn planting [8]. Yu et al. proposed multi-

layers framework to increase the performance of k -means algorithm on the dataset with outliers and noisy values [9]. In addition, genetic algorithm is used to obtain the optimal clustering results. Zhu et al. proposed a grid k -means algorithm to improve the clustering accuracy and stability and validated its performance on the dataset with the noise points [10]. Cuomo et al. used parallel techniques to reduce the computation cost of k -means algorithm for the large data analytic problem and provides solutions for the problems of GPU space limitation and host-device data transfer time [11]. k -means algorithm clusters data according to their spatial distance, resulting in it being difficult to ensure that the data in the same cluster have a similar or same relationship between the response of interest and input variables. Thus, it is necessary to develop a new clustering method under the framework of k -means algorithm.

In recent years, an interpolation model, Kriging method, has been widely used to model the relationship between the response and input variables of the measured data of the real-world systems. For example, Echard et al. assessed the failure probabilities of an engineering system using the importance sampling method and Kriging method, which has been successfully used in the reliability analysis of engineering systems [12]. Keshtegar et al. used Kriging method to estimate the solar radiation based on the meteorological data [13]. Wojciech proposed a digital terrain estimation method based on Kriging method, in which a neighbor points selection method is designed to accelerate the training speed Kriging method [14]. Belkhiri et al. estimate the groundwater quality for drinking purposes using Kriging method [15]. The results indicate that the Kriging model with electrical conductivity as co-variable produces the best performance compared with the other Kriging models. From the above works, it can be seen that Kriging method can effectively learn the relationship between the response of interest and input variables from the measured data. Thus, a k -interpolation model clustering algorithm is proposed based on Kriging method under the framework of k -means algorithm in this work, aims to partition data according to the relationship between the response of interest and input variables. Kriging method is used to evaluate the relationship between the response and input variables. For each datum, the estimation errors of the interpolation models of the clusters are used to decide its assignment. An optimization strategy is designed to obtain the clustering results. Finally, the performance of the proposed algorithm is validated through several synthetic and real-world datasets. The remainder of this work is organized as follows. The proposed algorithm

*Corresponding Author.

including the background of k-means algorithm and Kriging method is introduced in Section 2. The synthetic datasets and engineering datasets are used to test and compare the performance of the proposed algorithm with the conventional clustering algorithms in Sections 3 and 4. The conclusions are provided in Section 5.

II. LITERATURE REVIEW

In recent years, several data clustering algorithms have been proposed to partition data according to their relationship between the response of interest and input. Peng et al. [16] introduced ridge regression to evaluate the relationship of two-dimensional data in their clustering. Chen et al. [17] used the least square method to evaluate the features of data, and then applied fuzzy c-means algorithm to cluster them. However, only the linear relationship is considered in the above methods. To realize data clustering based on their nonlinear relationship between the response of interest and input, artificial neural networks and Gaussian process regression have been used to replace linear models. For example, Blažič et al. [18] used artificial neural networks to evaluate the nonlinear regression relationship to identify the state of engineering systems. Fuhg et al. [19] applied Gaussian process regression to evaluate the relationship among attributes to partition data according to their variation ranges. Fang et al. [20] used artificial neural networks to evaluate the relationship among data attributes to cluster the in-situ data of a tunnel boring machine.

III. PROPOSED METHOD

A. K-means Algorithm

k-means algorithm is developed in the area of signal processing, which aims to partition n data into k clusters in which each datum belongs to the cluster with the nearest mean (the prototype of the cluster). Generally, the clustering process of k-means algorithm can be subdivided into two stages: assignment step and update step as follows.

Assignment step: each datum is assigned to the cluster with the nearest prototype as follows

$$S_i^t = \{datum_p: \|d_{ip}\|^2 \leq \|d_{jp}\|^2 \forall j, 1 \leq j \leq k\} \quad (1)$$

where d represents the distance between the datum and the mean (Euclidean distance is usually used), and $datum_p$ is assigned to exactly one S_i^t .

Update step: the mean (prototype) of each cluster is recalculated as follows.

$$m_i^{t+1} = \frac{1}{|S_i^t|} \sum_{datum_j \in S_i^t} datum_j \quad (2)$$

The iterations are carried out until the assignments no longer change.

B. Kriging Method

In Kriging method (KRG), the following model is used to model the outputs at the samples:

$$Y(x) = f^T(x)\beta + Z(x) \quad (3)$$

where $Y(x)$ is the function of interest, $f = [f_1(x), f_2(x), \dots, f_m(x)]^T$ is the basis functions, and

$\beta = [\beta_1, \beta_2, \dots, \beta_m]^T$ is the corresponding coefficient vector. $Z(x)$ is a Gaussian stationary process with zero mean and covariance.

$$\text{Cov}(x_i, x_j) = \sigma^2 R(\theta, x_i, x_j) \quad i, j = 1, 2, \dots, n \quad (4)$$

where σ^2 is the process variance, $R(\theta, x_i, x_j)$ is the correlation function of the stochastic process, θ is the hyper-parameters of $R(\theta, x_i, x_j)$, and n is the sample number, The maximum likelihood method is used to optimize θ , where the likelihood function is expressed as follows:

$$L = (2\pi\sigma^2)^{-\frac{n}{2}} |R|^{-\frac{1}{2}} \exp \left[-\frac{1}{2\sigma^2} (Y - F\beta)^T R^{-1} (Y - F\beta) \right] \quad (5)$$

where R is the correlation matrix and F is a vector including the value of $f(x)$. β and σ^2 are estimated through the least-square method as follows.

$$\hat{\beta} = (F^T R^{-1} F)^{-1} F^T R^{-1} Y \quad (6)$$

$$\hat{\sigma}^2 = \frac{1}{n} (Y - F\hat{\beta})^T R^{-1} (Y - F\hat{\beta}) \quad (7)$$

By taking the logarithm of Eq. (5) with the imposed σ^2 value and multiplying by -1, the maximum problem to obtain the optimal θ is revised as

$$\text{minimize } \frac{1}{2} \ln(|R|) + \frac{n}{2} \ln(|\sigma^2|) \quad (8)$$

The prediction of Kriging method for a new sample x^* is

$$y(x^*) = f^T(x^*)\beta + r^T R^{-1} (Y - F\hat{\beta}) \quad (9)$$

where $r = [R(\theta, x_1, x^*), R(\theta, x_2, x^*), \dots, R(\theta, x_n, x^*)]$.

C. The Proposed Algorithm

A k-interpolation model clustering algorithm is proposed based on Kriging method in this section. From Eq. (1), it can be known that the distance d should involve the relationship between the response of interest and input variables, if we want to cluster the data based on the relationship. In this work, Kriging method is used to evaluate the relationship, and the estimated response of each datum can be obtained as flows.

$$\widehat{y}_{p,i} = \text{KRG}(x_p)_i \quad (10)$$

where $\widehat{y}_{p,i}$ is the estimated response of k -th datum of i -th cluster, x_p is the vector of input variables. The distance d is defined as follows.

$$d_{ip} = |y_p - \widehat{y}_{p,i}| \quad (11)$$

Similar to k-means algorithm, the clustering process of the proposed algorithm (named k-IM) is summarized as follows.

Step 1. Set the clustering number c ;

Step 2. Generate the assignment of the data randomly;

Step 3. Construct KRG model of i -th cluster based on the data contained in the cluster;

Step 4. Using the obtained KRG models to get the responses of all the data and creating the responses matrix $Y_{n \times c}$;

Step 5. Assigning each datum to the cluster using Eq. (1) and Eq. (11).

Step 6. If any stop conditions are satisfied, the procedure is stopped, and the current assignment results are considered as the final clustering results, otherwise, return to Step 3.

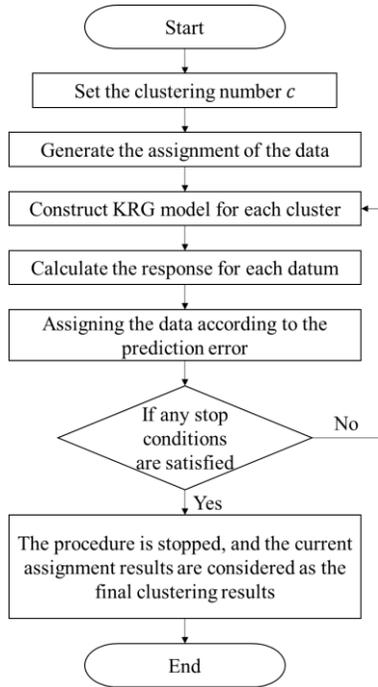


Fig. 1. Flow-chart of the Proposed Algorithm.

IV. EXPERIMENTS ON SYNTHETIC DATASETS

In this section, the synthetic datasets are used to validate the proposed algorithm. For each dataset, the data of each cluster is generated first and combined as the final dataset. The Latin hypercube sampling method is used to generate the input variables, and then the corresponding responses are calculated through the setting relationship between the response and input variables. The naming of the dataset is based on its sample number and cluster number. For example, N400C2 means that the dataset has 400 samples and two clusters. The proposed algorithm is compared with three popular clustering methods, *k*-means algorithm (KM), fuzzy *c*-means algorithm (FCM), and Gaussian mixture model (GMM). The clustering performance is evaluated through the following indexes.

1) Misclassification rate (MS):

$$MS = \frac{N_{error}}{N_{total}} \quad (12)$$

where N_{error} is the number of misclassified data; N_{total} is the total number of data. The lower *MS*, the higher cluster validity.

2) Adjusted rand index (ARI) [21]: Given a set S of n elements, and two partitions of these elements, namely $X = \{X_1, X_2, \dots, X_s\}$ and $Y = \{Y_1, Y_2, \dots, Y_s\}$, the overlap between X and Y can be summarized in a contingency table $[n_{ij}]$ where each entry n_{ij} denotes the number of objects in common between X_i and Y_j : $n_{ij} = |X_i \cap Y_j|$ as shown in Table I. Adjusted rand index is defined as follows:

$$ARI = \frac{\sum_{i,j} \binom{n_{ij}}{2} - \left[\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2} \right] / \binom{n}{2}}{\frac{1}{2} \left[\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2} \right] - \left[\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2} \right] / \binom{n}{2}} \quad (13)$$

The closer *ARI* to 1, the higher cluster validity.

TABLE I. CONTINGENCY TABLE

$X \setminus Y$	Y_1	Y_2	...	Y_s	Sums
X_1	n_{11}	n_{12}	...	n_{1s}	a_1
X_2	n_{21}	n_{22}	...	n_{2s}	a_2
...
X_s	n_{r1}	n_{r2}	...	n_{rs}	a_r
Sums	b_1	b_2		b_s	

3) Normalized mutual information (NMI) [22]:

$$NMI(X, Y) = \frac{I(X, Y)}{\sqrt{H(X)H(Y)}} \quad (14)$$

where $I(\cdot)$ is the mutual information metric and $H(\cdot)$ is the entropy metric. The closer *NMI* to 1, the higher cluster validity.

A. Effect of Sample Number

In this section, four synthetic datasets are used to study the effect of sample number on the performance of the *k*-IM algorithm. In each dataset, there are two clusters, and each cluster has the following relationship between the response and input.

$$\text{Cluster 1: } y = (6x - 2)^2 * \sin(12x - 4)$$

$$\text{Cluster 2: } y = 0.6(6x - 2)^2 * \sin(12x - 4) + 12(x - 0.5) + 6$$

where $x \in [0,1]$. For each synthetic dataset, one cluster has 150, 200, 250, 300 samples, respectively. Thus, the four synthetic datasets are denoted as N100A2C2, N200A2C2, N300A2C2, N400A2C2, respectively. The obtained N400A2C2 dataset is shown in Fig. 1. From Fig. 2, it can be seen that the samples of the two clusters are distributed similarly, but the relationship between the response of interest and input is different. The 30 times experiments are conducted for each dataset. The average experimental results are shown in Tables II to IV.

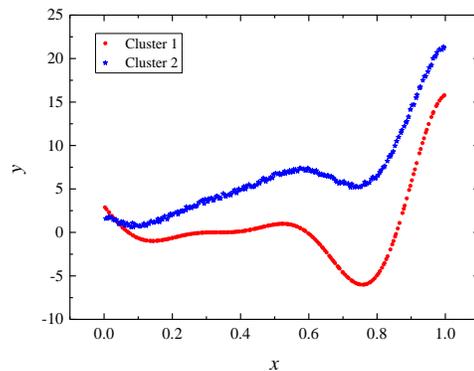


Fig. 2. N400C2 Dataset.

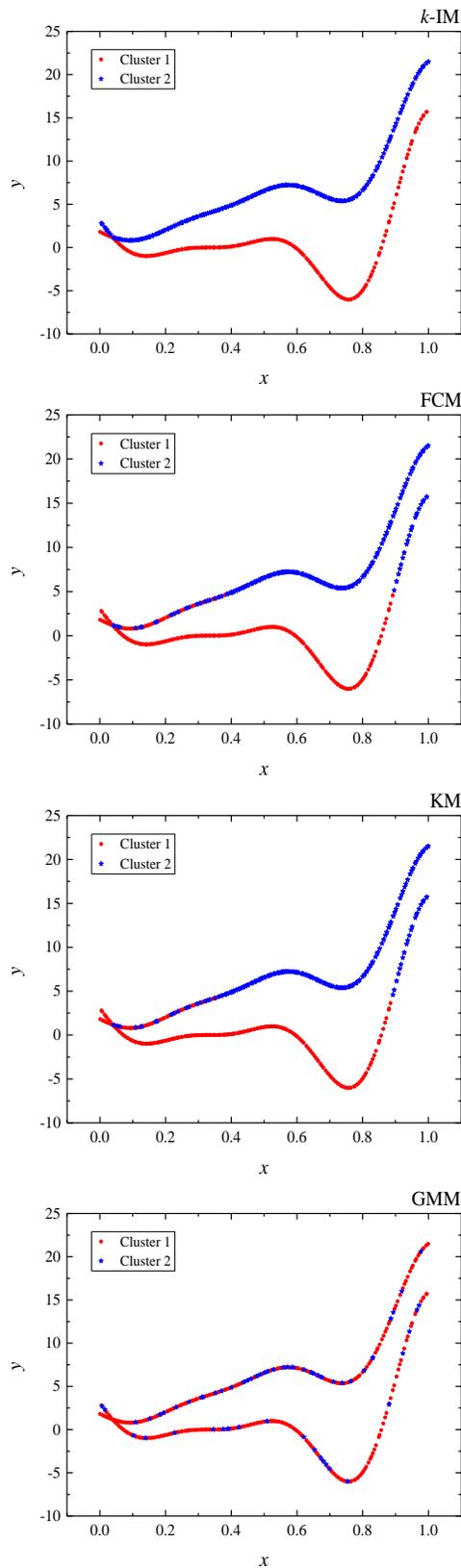


Fig. 3. Clustering Results Comparison of N400C2 Dataset.

From these tables, it can be seen that the k -IM algorithm produces much better results than the FCM, KM, and GMM

algorithms. The mean misclassification rate of the proposed algorithm is less than 0.03, which is much smaller than those of FCM, KM, and GMM, indicating the k -IM algorithm is able to accurately cluster the synthetic datasets. To further compare the performance of the clustering algorithms, the clustering results of N400C2 dataset of one experiment are shown in Fig. 3. From this figure, it can be seen that the proposed algorithm clusters the data based on the relationship between the output and input. The FCM and KM algorithms cluster the data according to their spatial distribution. It is noted that the GMM algorithm assigns most data to one cluster. The reason is mainly that it clusters data with the assumption that the data obey a Gaussian mixed distribution. The assumption cannot be stratified for N400A2C2 dataset. Thus, the clustering results of the GMM algorithm are much worse than the other algorithms. From the experimental results shown in Tables II to IV, it is observed that the sample number has an effect on the proposed algorithm. With the sample number increasing from 300 to 600, the MS of the proposed algorithm first decreases to 0.013 and then increases to 0.019. Similar results can be found in the indexes ARI and NMI. The reason is explained as follows. As the sample number increases, more samples can be utilized to construct the KRG models, which mean that the relationship between the output and input can be evaluated more accurately. The performance of the k -IM algorithm increases with the increase in the sample number. However, the KRG model tends to be overfitting when the sample number is too large. Thus, the performance of the k -IM algorithm decreases with the sample number increasing from 500 to 600. The proposed algorithm produces competitive clustering results for the datasets with different sample numbers tested in this section.

TABLE II. THE EXPERIMENTAL RESULTS OF MS

Dataset	k -IM	FCM	KM	GMM
N300C2	0.028	0.247	0.370	0.439
N400C2	0.027	0.248	0.391	0.467
N500C2	0.013	0.246	0.381	0.464
N600C2	0.019	0.246	0.339	0.485

TABLE III. THE EXPERIMENTAL RESULTS OF ARI

Dataset	k -IM	FCM	KM	GMM
N300C2	0.894	0.253	0.107	0.099
N400C2	0.894	0.251	0.084	0.080
N500C2	0.951	0.255	0.095	0.076
N600C2	0.926	0.256	0.146	0.069

TABLE IV. THE EXPERIMENTAL RESULTS OF NMI

Dataset	k -IM	FCM	KM	GMM
N300C2	0.833	0.215	0.096	0.094
N400C2	0.839	0.214	0.077	0.067
N500C2	0.923	0.216	0.085	0.063
N600C2	0.883	0.216	0.126	0.068

B. Effect of Cluster Number

Three datasets with different cluster numbers are used to test the effect of clustering number on the proposed algorithm, as shown in Table V. It is observed that the clusters of each dataset have similar but different relationships between the response and input variables. For each dataset, 30 times experiments are conducted. The average clustering results are shown in Fig. 4.

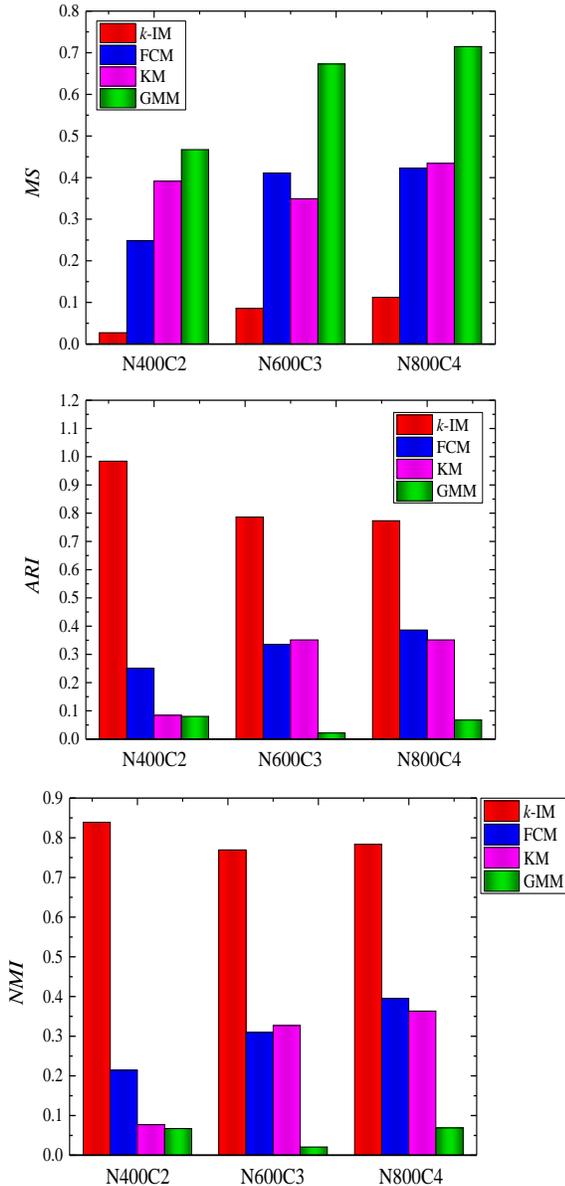


Fig. 4. Clustering Results of N400C2, N600C3 and N800C4 Datasets.

From Fig. 3, it is observed that the k-IM algorithm is still able to produce the best results among the tested four algorithms. The highest misclassification rate of the k-IM algorithm is around 0.10, which is much smaller than the conventional FCM, KM, and GMM algorithms. The index ARI of the k-IM algorithm is higher than 0.80 for all three datasets.

The index NMI is around 0.80, which is higher than the other clustering algorithms as well. It is noted that the misclassification rate of the GMM algorithm is higher than 0.50 for N600C3 and N800C4 datasets. The reason is that the GMM algorithm clusters almost all the data into one class, which means that most data are misclassified. Thus, the MS is higher than 0.50. With the cluster number increasing, the performance of the k-IM algorithm decreases, but it is still much better than the other popular clustering algorithms. The proposed algorithm can produce competitive clustering results when clustering the dataset with different cluster numbers tested in this section.

C. Effect of Noise

The measured data of real-world systems usually have noise. N400A2C2 dataset is used to test the performance of the k-IM algorithm on the noise. The synthetic datasets are generated as follows. For each cluster, the input variables are generated. The response is calculated according to the set function. For each datum, a random value is generated according to the set interval as shown in Table VI and added to the response. The average clustering results of 30 times experiments are shown in Tables VI to VIII.

TABLE V. EFFECT OF SAMPLE ON THE CLUSTERING PERFORMANCE (MS)

Dataset	Relationship
N400C2	$y = (6x - 2)^2 * \sin(12x - 4)$
	$y = 0.6(6x - 2)^2 * \sin(12x - 4) + 12(x - 0.5) + 6$
N600C3	$y = (6x - 2)^2 * \sin(12x - 4)$
	$y = (6x - 2)^2 * \sin(12x - 4) + 6(x - 0.5)$
N800C4	$y = (6x - 2)^2 * \sin(12x - 4)$
	$y = (6x - 2)^2 * \sin(12x - 4) + 6(x - 0.5)$
	$y = (6x - 2)^2 + 10$
	$y = 0.6(6x - 2)^2 * \sin(12x - 4) + 12(x - 0.5) + 6$

TABLE VI. EFFECT OF NOISE ON THE CLUSTERING PERFORMANCE (MS)

Noise	k-IM	FCM	KM	GMM
-	0.027	0.248	0.391	0.467
[-0.25,0.25]	0.029	0.246	0.404	0.432
[-0.50,0.50]	0.038	0.247	0.334	0.393
[-0.75,0.75]	0.040	0.248	0.390	0.413
[-1.00,1.00]	0.044	0.252	0.360	0.397

TABLE VII. EFFECT OF NOISE ON THE CLUSTERING PERFORMANCE (ARI)

Noise	k-IM	FCM	KM	GMM
-	0.894	0.251	0.084	0.080
[-0.25,0.25]	0.887	0.256	0.111	0.072
[-0.50,0.50]	0.854	0.255	0.173	0.116
[-0.75,0.75]	0.846	0.252	0.122	0.094
[-1.00,1.00]	0.830	0.245	0.139	0.107

TABLE VIII. EFFECT OF NOISE ON THE CLUSTERING PERFORMANCE (NMI)

Noise	k-IM	FCM	KM	GMM
-	0.839	0.214	0.077	0.067
[-0.25,0.25]	0.819	0.216	0.096	0.060
[-0.50,0.50]	0.774	0.216	0.147	0.094
[-0.75,0.75]	0.764	0.214	0.105	0.081
[-1.00,1.00]	0.741	0.209	0.120	0.087

The performance of the k-IM algorithm is better than the other popular clustering algorithms even if the dataset has noise in the relationship between the response of interest and input variables. The MS of the k-IM algorithm is smaller than 0.05, which means that less than five percent of the data are misclassified. Similar results can be found in the experimental results of the performance index ARI and NMI. With the noise level increasing, the performance of the k-IM algorithm decreases. When the dataset has higher noise in the relationship between the response of interest and input variables, the Kriging method is more difficult to accurately evaluate the relationship. Thus, the performance of the proposed algorithm is worse when the noise level is higher. But, the MS of the k-IM algorithm is still smaller than 0.045. The k-IM algorithm can produce competitive clustering results for the datasets tested in this section.

V. EXPERIMENTS ON ENGINEERING DATASETS

In this section, three engineering datasets are used to further test the proposed algorithm. Since the classification information of the engineering datasets is unknown, the experiments are conducted as follows. For each engineering dataset, the dataset is first clustered into several subsets. For each subset, five cross-validation method is used to test whether the data in the same subset has a similar relationship between the response of interest and input variables. The subset is randomly divided into five parts, one part is selected as the testing data, and the remaining four parts are used as the training data. The experiments are conducted five times, and the average R-square of the five experiments is used to assess the consistency of the relationship of the subset. R-square is calculated as follows.

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (15)$$

where n is the number of the testing data, y_i is the real response, \hat{y}_i is the estimate response, and \bar{y} is the mean of the real responses. The closer R^2 to 1, the better performance.

A. Yacht Hydrodynamics Dataset

The yacht hydrodynamics dataset is first used. The dataset comes from a series of 308 experiments on the residuary resistance of sailing yachts [23]. Several input variables are considered, including the prismatic coefficient, longitudinal position of the center of buoyancy, length-displacement ratio, beam-draught ratio, length-beam ratio, and Froude number. The residuary resistance is evaluated through the per unit weight of displacement. The k-IM, FCM, KM, and GMM

algorithms are used to cluster the yacht hydrodynamics dataset into two subsets. And, five cross-validation methods are applied to each subset to test whether the data in the same subset has a similar relationship between the residuary resistance and input variables. The experiments are conducted 30 times, and the corresponding results are shown in Fig. 5. The average R^2 of the k-IM algorithm is higher than 0.98 for the obtained two clusters, which is much better than that of the FCM, KM, and GMM algorithms, indicating that the data of the same cluster obtained by the proposed algorithm have more similar relationship between the response of interest and input variables. The k-IM algorithm is able to cluster the yacht hydrodynamics dataset according to the relationship between the residuary resistance and input variables.

B. Bolt Tensioner Dataset

Bolt tensioner is a widely used tensioning tool in the assembly of large equipment such as nuclear power generators or the construction of large buildings [24]. It is an annular jack that rises up through hydraulic pressure. The bolt tensioner dataset recorded the data from 40 simulations, including the maximum stress at the piston of the bolt tension and the corresponding structural parameters with the same hydraulic pressure. In the experiment, the cluster number is set two as well, and the k-IM, FCM, KM and GMM algorithms are used to cluster the dataset. Based on the clustering results of each clustering algorithm, five cross-validation methods are to test whether the data in the same cluster has a similar relationship between the maximum stress and structural parameters. Fig. 5 shows the experimental results. It is noted that the GMM algorithm cannot provide clustering results since the covariance matrix is ill. From Fig. 6, it can be seen that the average R^2 s of the k-IM algorithm is the highest among the tested four clustering algorithms. The k-IM algorithm is able to cluster the bolt tensioner data such that the data in the same cluster have a similar relationship between the maximum stress and structural parameters.

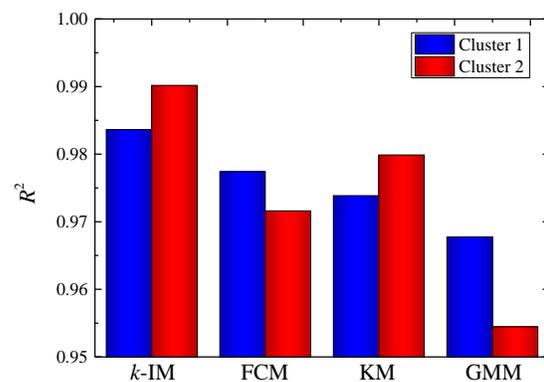


Fig. 5. Experimental Results of Yacht Hydrodynamics Dataset.

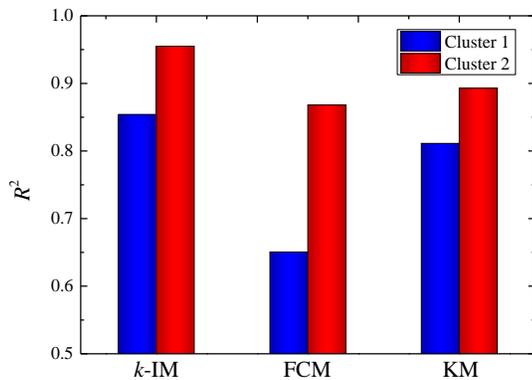


Fig. 6. Experimental Results of Bolt Tensioner Dataset.

VI. CONCLUSION

In this work, we proposed a k -interpolation model clustering algorithm (named k -IM) to cluster data according to the relationship between the response of interest and input variables. In the proposed algorithm, Kriging method is used to construct the interpolation models. For each datum, the estimation errors of the interpolation models of the clusters are used to decide its assignment. An optimization strategy is designed to obtain the clustering results under the framework of k -means algorithm. The effect of the sample number, cluster number, and noise level on the k -IM algorithm is studied through several synthetic datasets. The results indicate that the k -IM algorithm in this paper can provide competitive clustering results. Two engineering datasets are further to test the performance of the k -IM algorithm as well, and the experimental results show that the k -IM algorithm is able to cluster the data such that the data in the same part have a similar relationship between the response of interest and input variables.

ACKNOWLEDGMENT

The authors would like to thank the editors' and reviewers' work on this paper.

REFERENCES

- [1] I. Škrjanc, J. Iglesias, A. Sanchis, D. Leite, E. Lughofer, and F. Gomide, "Evolving fuzzy and neuro-fuzzy approaches in clustering, regression, identification, and classification: A survey," *Information Sciences*, vol. 490, pp. 344-368, 2019.
- [2] A. Dubey and A. Rasool, "Clustering-based hybrid approach for multivariate missing data imputation," *International Journal of Advanced Computer Science and Applications*, vol. 11, pp. 710-714, 2020.
- [3] X. Song, M. Shi, J. Wu, and W. Sun, "A new fuzzy c-means clustering-based time series segmentation approach and its application on tunnel boring machine analysis," *Mechanical Systems and Signal Processing*, vol. 133, pp. 106279, 2019.
- [4] R. Xu and D. Wunsch, "Survey of clustering algorithms," *IEEE Transactions on neural networks*, vol. 16, pp. 645-678, 2005.
- [5] M. Shi, T. Zhang, L. Zhang, W. Sun, and X. Song, "A fuzzy c-means algorithm based on the relationship among attributes of data and its

- application in tunnel boring machinem," *Knowledge-Based Systems*, vol. 191, pp. 105229, 2020.
- [6] J. Diaz-Rozo, C. Bielza, and P. Larrañaga, "Clustering of data streams with dynamic gaussian mixture models: an IoT application in industrial processes," *IEEE Internet of Things Journal*, vol. 5, pp. 3533-3547, 2018.
- [7] A. Shubair, and A. Al-Nassiri, "kEFCM: kNN-based dynamic evolving fuzzy clustering method," *Proc. IJACSA*. vol. 6, pp. 5-13, 2015.
- [8] A. Aldino, D. Darwis, A. Prastowo, and C. Sujana, "Implementation of K-means algorithm for clustering corn planting feasibility area in south lampung regency," *Journal of Physics: Conference Series*. vol. 1751, pp. 012038, 2021.
- [9] S. Yu, S. Chu, C. Wang, Y. Chan, and T. Chang, "Two improved k-means algorithms," *Applied Soft Computing*, vol. 68, pp. 747-755, 2018.
- [10] E. Zhu, Y. Zhang, P. Wen, and F. Liu, "Fast and stable clustering analysis based on Grid-mapping K-means algorithm and new clustering validity index," *Neurocomputing*, vol. 363, pp. 149-170, 2019.
- [11] S. Cuomo, V. Angelis, G. Farina, L. Marcellino, and G. Toraldo, "A GPU-accelerated parallel K-means algorithm," *Computers & Electrical Engineering*, vol. 75, pp. 262-274, 2019.
- [12] B. Echard, N. Gayton, M. Lemaire, and N. Relun, "A combined importance sampling and Kriging reliability method for small failure probabilities with time-demanding numerical models," *Reliability Engineering & System Safety*, vol. 111, pp. 232-240, 2013.
- [13] B. Keshtegar, C. Mert, and O. Kisi, "Comparison of four heuristic regression techniques in solar radiation modeling: Kriging method vs RSM, MARS and M5 model tree," *Renewable and sustainable energy reviews*, vol. 81, pp. 330-341, 2018.
- [14] M. Wojciech, "Kriging method optimization for the process of DTM creation based on huge data sets obtained from MBESs," *Geosciences*, vol. 8, pp. 433, 2018.
- [15] L. Belkhir, A. Tiri, and L. Mouni, "Spatial distribution of the groundwater quality using kriging and Co-kriging interpolations," *Groundwater for Sustainable Development*, vol. 11, pp. 100473, 2020.
- [16] C. Peng, Q. Zhang, Z. Kang, C. Chen, and Q. Cheng, "Kernel two-dimensional ridge regression for subspace clustering", *Pattern Recognition*, vol.113, pp.107749, 2021.
- [17] Y. Chen and Z. Yi, "Locality-constrained least squares regression for subspace clustering," *Knowledge-Based Systems*, vol.163, pp.51-56, 2019.
- [18] S. Blažič and I. Škrjanc, "Hybrid system identification by incremental fuzzy c-regression clustering," In *2020 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, pp.1-7, 2020.
- [19] J. N. Fuhg and A. Fau, "A classification-pursuing adaptive approach for Gaussian process regression on unlabeled data," *Mechanical Systems and Signal Processing*, vol.162, pp.107976, 2022.
- [20] J. Fang, X. Song, N. Yao, and M. Shi, "Application of FCM Algorithm combined with artificial neural network in TBM operation data," *Computer Modeling in Engineering & Sciences*, vol.126, pp.397-417, 2021.
- [21] K.Y. Yeung and W. L. Ruzzo, "Details of the adjusted rand index and clustering algorithms, supplement to the paper an empirical study on principal component analysis for clustering gene expression data," *Bioinformatics*, vol. 17, pp. 763-774, 2001.
- [22] P.A. Estévez, M. Tesmer, C.A. Perez, and J.M. Zurada, "Normalized mutual information feature selection," *IEEE Transactions on neural networks*, vol. 20, pp. 189-201, 2009.
- [23] I. Ortigosa, R. Lopez, and J. Garcia, "A neural networks approach to residuary resistance of sailing yachts prediction," In *Proceedings of the international conference on marine engineering MARINE*, vol. 2007, pp. 250, 2007.
- [24] J. Liu, Z. Zhang, M. Wang, J. Wang, and S. He, "Main bolt tensioner for pressure vessel of 10 MW high temperature reactor," *Nuclear Power Engineering*, vol. 21, pp. 503-506, 2000.

A Food Waste Mobile Gamified Application Design Model using UX Agile Approach in Malaysia

Nooralisa Mohd Tuah¹, Siti Khadijah Abd. Ghani², Suryani Darham³, Suaini Sura⁴

Faculty of Computing & Informatics, Universiti Malaysia Sabah, Labuan International Campus, 87008 Labuan F.T, Malaysia^{1,4}
Biodiversity Management Division, Ministry of Energy and Natural Resources, Presint 4, 62574 Putrajaya, Malaysia²
Institute of Tropical Agriculture and Food Security, Universiti Putra Malaysia, 43400 Serdang, Selangor, Malaysia³

Abstract—Food waste is a significant worldwide issue in landfill management. Due to improper implementation, technology applications related to food waste collection and its management system are still lacking in practice. The available applications have yet to address the issue of food waste management. Constructing an interactive mobile application is necessary for managing food waste collection for the decomposition process using Black Soldier Fly (BSF) treatment. Furthermore, as the mobile application requires participation from various user backgrounds, maintaining user involvement has become a priority. Gamification has emerged as one of the approaches that might favourably affect individual engagement behaviour. A comprehensive game element design is required where it focuses on how gamification can influence user engagement. This study aims to model the food waste gamified mobile application design to benefit Malaysia's decomposition ecosystem. It includes gamification, management features, and data visualization for reporting and will involve users from households, businesses, and the BSF farm. This paper presents the modelling process of a new mobile application design for this concept of study. The UX agile approach was used in gathering and designing the application requirements as it allows for active participation from all stakeholders. The result shows that the experts agree on the application design. This research will indirectly benefit the BSF industry in Malaysia, and it will have a significant impact on gamification, user experience, and food waste management in the direction of a sustainable environment.

Keywords—Avatar; food waste disposal; mobile apps; gamification; data visualization; black soldier fly

I. INTRODUCTION

Information system focuses on providing solutions to business processes for better operation management, information, and supporting the decision-making process. Recent technological advancements have transformed the way business processes are carried out. In this context of the study, the technology transformation includes the business process of collecting food waste from sources (households and businesses) and sending it to farms for treatment using black soldier fly (BSF). This treatment is one of the methods used to keep organic waste out of landfills [1]. With BSF as a treatment agent, food waste disposal is environmentally safe and cost-effective [1 – 2]. It is known that BSF decomposes food waste into animal feed and compost. Thus, the quality of food waste is critical for BSF production. As a result, food waste preparation should adhere to the guidelines provided. Therefore, food waste producers should be cultivated with proper, clean, and safe disposal. It is necessary to simplify the

procedure of food waste disposal for them to actively participate in keeping the surrounding environment safe [3]. This procedure could involve everything from waste preparation to waste collection and delivery to the farm for further processing.

As smartphones become more prevalent in daily life, using them to support food waste processes ranging from waste cultivation to collection will allow individuals and society to contribute to environmental sustainability. However, in order to contribute, they must be driven or motivated. Current food waste collection, disposal, and decomposition systems or mobile applications on the market are not related to collection, disposal, or decomposition. They are merely about raising awareness about food waste and recycling [3 – 5]. Although technology intervention in food waste operations may be attracting much attention worldwide, it is unlikely to be internalized for use by other landfills, particularly those commercially produced for the Malaysian market. This paper argues that it is deemed necessary to have an application that can stimulate community involvement and an application that can assist landfills in conducting their business operations, particularly in the case of the Malaysian context of studies. Gamification offers a game-like experience to a serious or monotonous type of work. This experience will make the users interact with the system not because they are forced to but because they want to use it and be a part of a fun and appealing system. As gamification aims to motivate users, its application may benefit food waste management. Furthermore, because the national aspiration is toward digitization, there seems to be a significant effort to improve the organization's information system management in all aspects of the operation.

This paper presents the development and modelling of a smart mobile gamified application for managing local food waste, a case study in the Malaysian context. Using BSF, the technology will support food waste requests and collection business processes for later decomposition arrangements. This technology is made available through mobile apps that include gamification, waste management features, and data visualization for reporting. The application can help to support the government effort to preserve the environment, especially the food waste disposal and ecosystem of BSF. For that purpose, the groundwork of the system application design and development are described, the results from the design usability testing are presented, and this design later be modelled as a guideline for similar system implementation. Given the current literature review, it shows that this research

would be a significant digital application if implemented in the context of Malaysia.

The following is how this paper is structured: First, as an introduction to the study, this paper reviews work related to food waste and BSF in Malaysia, food waste technology adaptation, and the use of gamification in the related application. Second, it describes the study's materials and methods. The description includes the development methodology, the mobile application's requirements and design, as well as design validation. The third section contains a discussion of the results of the design usability test. A model for such an application is created as a result of the design activities and is presented in section four. Finally, concluding remarks and recommendations for future work are presented in the final section.

II. RELATED WORK

A. The Model Development Process

Historically, most countries, developed or developing, have relied on landfills as final waste disposal sites [6 – 7]. Today, many developed countries have made other options within the waste hierarchy mandatory to reduce landfilling. Developing countries, particularly impoverished urban areas, rely on unsustainable solid waste management, primarily dumpsites. Over 90% of waste generated in low-income countries is either openly burned or disposed of in landfills [8]. Dumpsites receive all types of waste, including organic waste, and exposure to rainfall causes the production of leachate and methane. It contributes to the problem of climate change. Unknowingly, food waste is one of the wastes that has significantly contributed to the global dumpsite problem.

Malaysia had relatively high percentages of food waste from restaurants and households [3, 9]. According to Abd Ghafar [9], in 2017, restaurants contributed approximately 15,000 tonnes of food waste per day, with households contributing approximately USD 50 per month. In 2021, daily food waste will reach 38,000 tonnes, with household waste increasing by around USD 90 per month [10]. The local authorities are now dealing with a critical situation. One alternative to managing food waste is Black Soldier Fly (BSF). BSF will consume the food waste feed and compost the waste [1 – 2]. It is critical to preserve the ecosystem to create a sustainable environment. As a result, various approaches from various perspectives have been explored to support the effort. One of them is through the innovation of smart technology in assisting the food waste disposal processes for the benefit of BSF.

Food waste disposal management in the literature is centred on food recycling [3 – 4], landfill issues [4], and educational awareness [3, 5]. However, none of this is used commercially in Malaysia. Little is known about information technology that may impact food waste disposal processes and BSF production. High amounts of food waste are required to increase BSF production. Furthermore, the food waste disposal ecosystem should include households, retailers, agricultural industries, the local community, local governments, and the BSF industry. However, due to the limited venue or platform to contribute, communities may find the effort is not encouraged by the local

authority or another party. Thus, different individuals' behaviour and perceptions of the importance of beneficial food waste disposal emerge [11]. Besides, the community's awareness of the processes may not be widespread. As a result, there is a lack of involvement and motivation from the communities (restaurant, farm, household) [11 - 12]. In this view, communities may fail to see and realize the urgency to properly dispose of their food waste.

B. Technological Approach to Food Waste Management

Various alternatives should be developed to ensure a smooth domestic food waste disposal process, including food waste readiness, collection, sending to the appropriate farm (validated by local authorities), and verifying food waste disposal compliance. Research by Varhana, Faliasthiunus & Ulfah [13] has recently completed the development of a management system specifically for BSF. They created a traceable waste system that can track, control, and manage waste problems following their authority's policies. They used a waste bank, which served as a waste recipient, distributor, and manager. BSF would be able to generate a high economic value through the waste bank. However, since this application is primarily concerned with the management of BSF production, local communities, businesses, and government involvement may be given less emphasis. Other food waste mobile apps that have been developed are more focused on raising awareness [5], the Zero Food Waste Act [14], and ways to reduce food waste through donation [15]. In Jain et al. [5], gamification is used in a mobile application to encourage students to avoid making messes and wasting food and develop positive behaviour in maintaining a good environment. However, these current inventions do not include management of the food waste disposal process that involves societies for a sustainable environment. Some incentive that encourages them to be a part of the process should be designed and implemented to increase community participation.

Gamification is a practical approach to information management [16]. Gamification is expected to provide a fun and engaging environment for users to complete assigned tasks [16]. Ulla Santti et al. [17] investigated the use of gamification in encouraging users to process and sort their food waste. According to the study's findings, the gamified application can be effectively used to support a change in consumer behaviour, particularly in cultivating food recycling among young adults. Research by Ali & Ahmed [18] created a mobile geo-located system to locate and pattern user-reported food waste disposal. Gamification was used by the application to encourage user engagement with the application. Ali & Ahmed contend that increasing food waste awareness goes beyond food waste reduction. Users must be motivated to participate in the process. All available applications, such as those in [5, 17–18], are concerned with encouraging users to reduce food waste. In this regard, the gamification application differs from the proposed work for this research. However, the use of a motivating element to justify gamification is similar.

The local authority also requires an intelligent system to manage their waste administration better [9, 14]. In this view, a local authority may need a smart mobile application to cater to food waste movements. The application should involve the local communities, businesses, and BSF farms. An application

that allows them to actively monitor the production of BSF livestock from the collection of food waste seems demandable. Such application will indirectly help to improve the production of BSF livestock.

III. MATERIALS AND METHODS

This section describes the modelling and verification process of developing the proposed mobile application. The description includes the methodology used in the study, the detailed modelling processes involved, and the materials used in the study. The requirement analysis and application design activities in modelling the interactive application for managing household food waste for BSF in Malaysia are presented in detail.

A. Agile UX Methodology

The modelling and verification of the mobile application development follow the agile UX methodology. The methodology was chosen due to the nature of the work that requires active involvement from a system developer, project management, and other system stakeholders. Obtaining collective opinions and approval will be difficult if meetings do not exist or are held infrequently [19]. The study by Akhbar et al. [20] shows that software companies in Malaysia have been opting to adopt an agile approach in their software development projects. This adoption is due to the company's agreement to support the global software development projects, outsourcing the tasks to other offshore companies. The approach seems to have encouraged the collaboration of ideas, staff skills, and company resources between the local and offshore companies. Besides, agile implementation in the context of IS projects in Malaysia was a successful approach in delivering a rapid system implementation [21].

The agile UX development lifecycle is a method variation that builds on the Scrum framework [22]. It is organized into two tracks; the agile team and the UX team. The hybrid methodology consists of several phases that are interconnected between the agile team and UX team during the requirements, design, and testing phases. This development lifecycle follows the agile UX framework proposed by Kieffer, A., Ghouti, A., & Macq, B. [19], as illustrated in Fig. 1.

The requirement analysis and designs of the application will be the core and critical elements for this context of the study. Because the process involves the developer, designer, researcher, and potential users (e.g., admin farm, household, restaurant, local authority representation), all stakeholders will validate the requirements, design, and feedback.

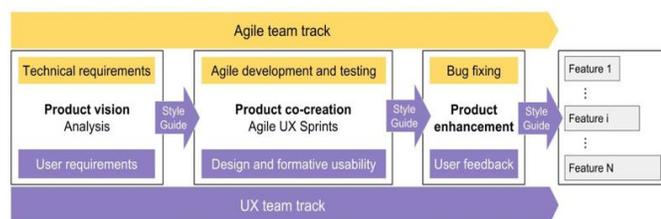


Fig. 1. Agile UX Development Lifecycle. Note. This Model was Adopted from Kieffer et al. [19].

B. Agile UX phases

Based on the methodology in Fig. 1, the application of each phase in this study is described as follows:

1) *Product vision*: This phase addresses the users and technical requirements. The UX team conducts user requirements, while the agile team obtains technical requirements. The UX team used personas (user stories) and task analysis to analyze the stakeholder's requirements. Meanwhile, technical requirements are gathered based on system development, team, and deployment requirements. The application's specific features are obtained at the end of this phase. Activities for user requirements are explained in the next section, requirement analysis.

2) *Product co-creation*: The team will design the application user interfaces and create a high-fidelity prototype for each feature during this phase. The designs are the initial features that have been proposed. The team will improve these designs regularly through the sprint activities. This phase also involves two groups, one from the agile team and one from the UX team. The agile team will create a product backlog for each feature and update it as the feature progresses for each sprint.

Meanwhile, the UX team will validate the design by conducting usability testing on each design and prototype. They will then decide what design changes are required in the next sprint. The sprint iterated until each design feature was released. Once released, the high-fidelity prototype will be transferred to development for final product deployment. This phase's research activity is to report on the study's initial design and design verification sections.

3) *Product enhancement*: The agile team will concentrate on product enhancement, investigate the errors, and correct them as necessary. Meanwhile, the UX team will reanalyze key user feedback and confirm the current designs. In addition, the analysis will feed requirements into other features in an indirect manner.

C. Requirement Analysis

1) *Scrum meeting*: Two scrum meetings were conducted to determine the system's requirements. The first meeting is for the initial user story and requirement, and the second is to confirm the stories and requirements. The official meeting was conducted online due to the different geo-location and limitations during the pandemic situation. The session involves developers, a system consultant, the system owner, and a system user. Non-official meetings were also held, which were conducted via phone calls and group messages. The consensus of each member on the required features is gathered and synchronized with the user stories. The requirements were later transferred to the initial designs.

2) *Personas*: A user story is essential for understanding the roles involved in the system application and which features should be designed according to the roles. It aids in

clarifying the system's scope and clearly defining users' roles in the system. In this study, the system has three main user stories. The story is as follows.

- **Collector (rider):** A collector is a rider who collects waste from a pickup location and transports it to a designated landfill site. They must apply to be a collector in the application and wait for the admin approval. They will be issued a collector's ID once they have been approved. The application will be using Google Maps to show the location with a tracking destination to assist a rider in picking up the waste. Collectors need to scan a QR code at the pickup and sending locations and confirm the pickup items. Riders must log in to their page to see their dashboard on completed tasks, new collection tasks, uncompleted tasks, history tasks, and revenue collected from the tasks. The rider will be able to update their profile and see their achievements through the gamification elements on their page.
- **Household or restaurant:** A household/restaurant manager ensures that food waste is ready for collection. They must bag the waste according to the instructions and then be prepared with the description, number of bags, and total weight to be entered into the application. In the application, they can find instructions and information on what is and is not collectable food waste. The household must book the available date and time for collection in the application. Aside from that, they will be able to view the entire collection history and the following scheduled collection. The application also allows for scheduling and cancellation. Households can view their accomplishments and contributions to an eco-friendly environment on their profile page.
- **System Admin:** An administrator is a person in charge of the system. The administrator will be able to manage the users. After checking and verifying user profiles, they will manually assign the system's users' roles (collector, household, and restaurant). In addition to auto collection allocation, the admin will be able to assign a collector to collect waste manually. Admin can view and report on all transactions. The admin is also the person in charge of approving waste deliveries and processing payments to the collector.

D. Initial System Design

1) *Application function flow:* Following the requirement, the user story is used to define the system flow. There are three flows in total: 1) food waste pickup collection flow (Fig. 2a), 2) household/restaurant request for collection (Fig. 2b), and 3) Admin manage the system (Fig. 2c).

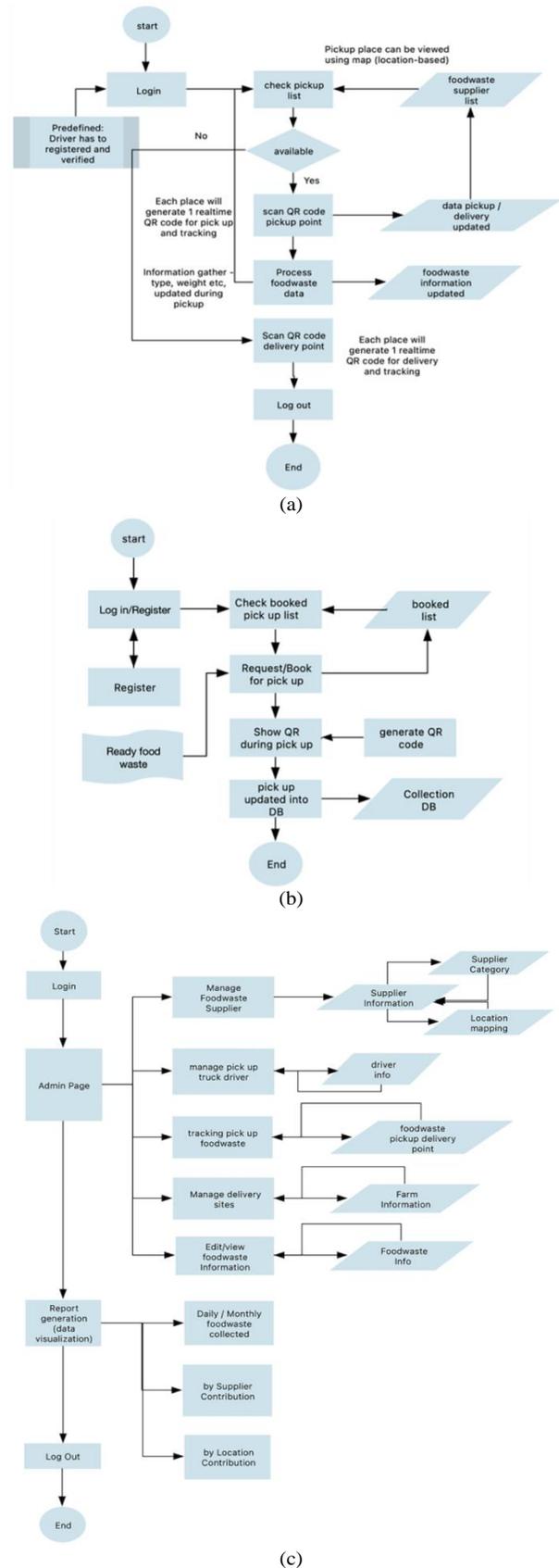


Fig. 2. (a) Food Waste Collection Flow Chart, (b) Requesting for Collection Flow Chart, (c) Admin Site Flow Chart.

2) *Application design:* The Figma application was used to create the high-fidelity prototype. Using Figma, it is possible to quickly visualize and present the application to users during the design sprints. The design comprises three main parts in line with the user story. The design includes the screen for signing up and login (see Fig. 3), the main screen for a collector - dashboard of a collection task, transaction history, maps direction for collection, user's profile – with gamification element, and QR code generation (Fig. 4a and 4b). It also includes the main screen for household and restaurant - collector profiles with gamification, transaction history, rewards, and QR code generation (Fig. 5a and 5b).

The gamification (as in Fig. 6a) included points, ranking, level, and rewards. Riders will be awarded points based on the total number of kilometres collected. When they reach specific points, such as 1km to 10,000km, they will advance to the intermediate level, 10,001km–50,000km, and 50,001km and above, they will go to the advanced level. The landfill's administrator can adjust the level. The rank element is determined by the rider's income from the collection. Each rider can convert their points into physical rewards based on the redemption offer made by the participating party.

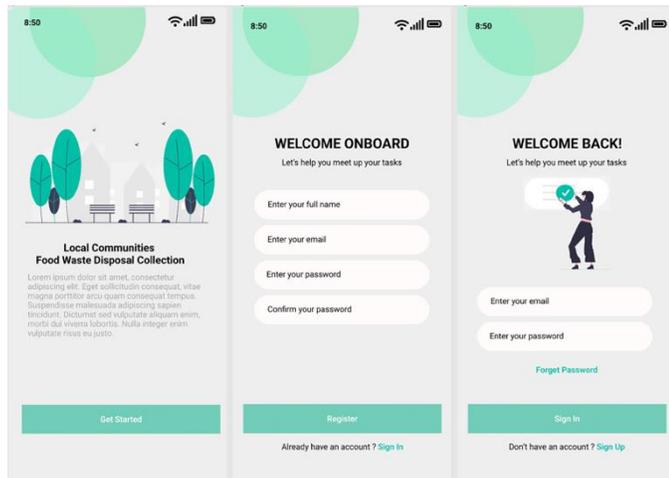
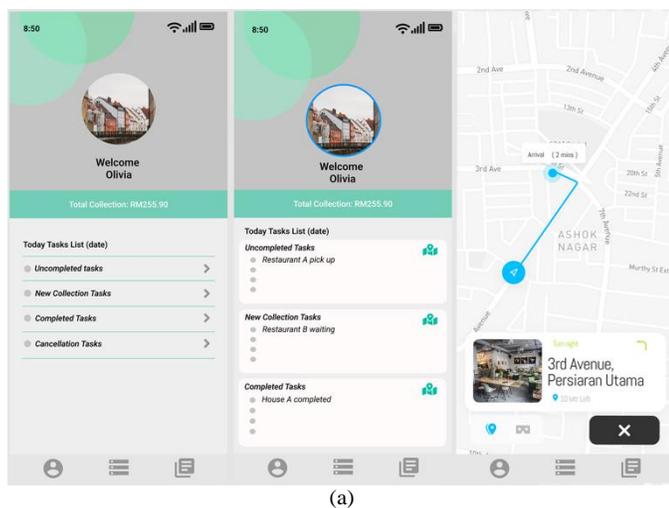
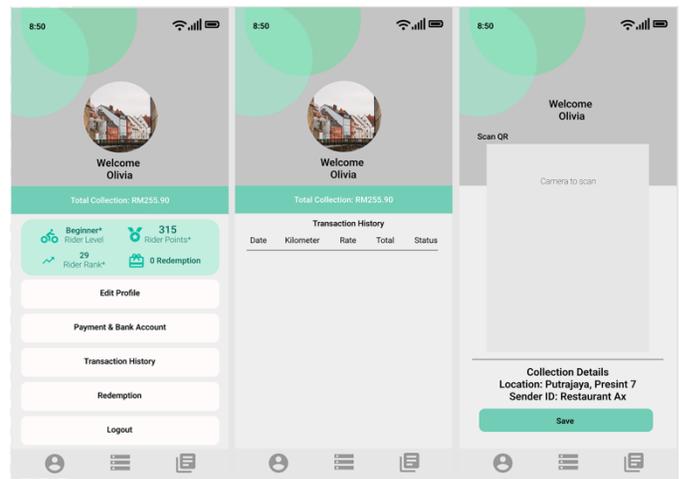


Fig. 3. The Main Screen, Sign Up and Login Screen.

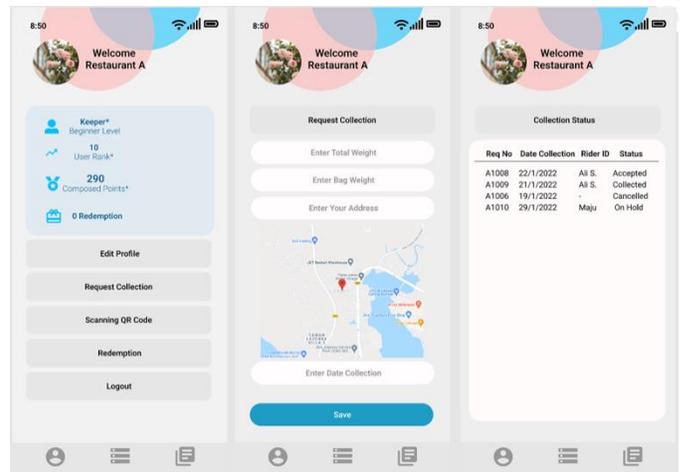


(a)

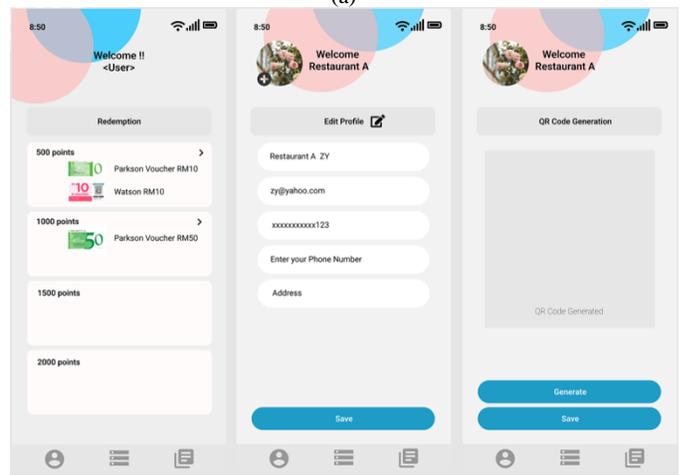


(b)

Fig. 4. (a) Collector's Screen – Dashboard, Task and Direction, (b) Collector's Screen – Profile and Transaction History.



(a)



(b)

Fig. 5. (a) Household/Restaurant Screen – Dashboard and Request, (b) Household/Restaurant Screen – Rewards and Profile.

Gamification elements in the restaurant or household function include points, ranking, level, and rewards (Fig. 6b). Like the rider, restaurants/households will be rewarded with

points based on the amount of waste they send to landfills. When they reach a certain kilogram weight, for example, between 1kg and 1,000kg, their level is changed to the keeper. If the total weight is between 1,001kg and 5,000kg, their level changes to a saviour, and if the total weighs more than 5,001kg, their level changes to a knight. The rank is determined by the total number of kg contributed. The points accumulated can be redeemed based on the redemption offer.

E. Design Validation

The designs were validated in the second phase of the Agile UX method. The validation was carried out to ensure that the design applications complied with the requirements and specifications of the users. Six participants were recruited for this purpose to provide feedback on the application's flow and their designs. The process was repeated several times until the final design was reached. The study was carried out via an online sprint meeting. Participants were asked to review the design, flow, and expected applicability when thoroughly applied.

1) *Participant and research design:* The selection of participants is based on their roles in the system and their responsibility in the system development. Generally, the roles involve collector, food waste supplier – households/restaurants, system owner, system developer, designer, and system consultant. These participants were also the key users of the system. In this manner, they were involved from the requirement phase to the design phase. They are also classified as someone being technologically savvy, as their daily activities involve the use of online systems and applications. The same participants were involved in assessing the proposed systems model design against design specification and system usability – ease of use in completing a specified task, consistency and standard design, and its practicality in implementation. The demographic of the participants is summarised in Table I.

TABLE I. PARTICIPANT'S DEMOGRAPHIC

Demographic	No
Gender	
Female	4
Male	2
Age	
25 - 29	1
30 - 34	0
35 - 39	1
40 - 44	4
Type of Users	
Key users (Non-system developers)	4
Non-key users (System developers)	2
Field of Expertise	
Key User	2
UX/UI researcher and developer	2
Landfill and dumpsites	2

2) *Application verification:* Before the study began, the researcher formally emailed each sprint meeting's invitation to the participants. The email was sent to obtain their consent and provide detailed information about the study. Following their consent, the online meeting details were emailed to them for confirmation. The meeting begins with a presentation of the system (status and features), usability testing, and a comments and discussion session. The session will typically last 30-45 minutes.

A list of tasks based on the system flow and design is provided for usability testing. The tasks are listed in the following Table II. For this testing, two materials were used – 1) the list of tasks to obtain usability problems raised during each sprint session, and 2) the types of user error measurement usability metric adopted from Kieffer et al. [19] (as shown in Table III) to indicate any usability problem found during the test.

TABLE II. LIST OF TASKS FOR USABILITY TESTING

Sprint (S)	Tasks	
Sprint 0:	T1	Sign up, log in, forget password
	T2	View all collection tasks
	T3	Select new task collection and view map for direction
	T4	Request collection (Restaurant)
Sprint 1:	T5	Edit profile, change a picture, edit payment bank info (both collectors and restaurant)
	T6	View transaction history and collection status (both collectors and restaurant)
	T7	Redemption function (both collectors and restaurant)
Sprint 2:	T8	QR code generation and scanning (both collectors and restaurants)
	T9	Review gamification elements (both collectors and restaurants)
	T10	Main menu review (both collectors and restaurant)

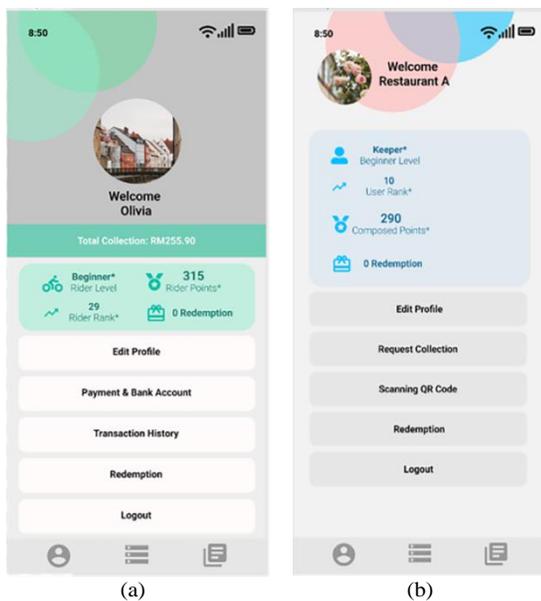


Fig. 6. (a) Rider's Gamification.(b) Household/Restaurant Gamification.

TABLE III. TYPE OF USER ERROR

Code	Type of Error
UE1	Behaviours that prevent task completion
UE2	Mistaken believes that a task is completed when it is not (or vice versa)
UE3	Misinterpretation of the content
UE4	Oversight of something that should be noticed
UE5	Expression of frustration by the participant
UE6	Participants remark about the possibility of improvement

During the online sprint meeting, participants will personally conduct the test of the design prototype according to the given tasks. Then, participants will note the types of errors they discovered for each system design prototype. For each error discovered, the severity of the error (low, medium, high) will be noted to determine what actions should be taken. Once finished, they will send back the error list to the researcher.

IV. RESULT AND DISCUSSION

1) *Usability testing*: Table IV summarizes the indications of usability errors. The authors discovered that participants misinterpreted the content of a few tasks: task 3, task 8, and task 9. One task (task 5) was mistakenly thought to be completed, while another task (tasks 5&9) was suggested for improvement. Overall, the errors were rated as having a low risk of seriousness. The low risk indicates that the issues raised are not likely to cause task failure, but they may have a minor impact on system efficiency or user satisfaction. Thus, feedback from participants was gathered through comments and discussion sessions for detailed explanation and to work out the proposed solution.

TABLE IV. SYSTEM DESIGN USABILITY ERRORS

Sprints	Sp0				Sp1			Sp2		
	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10
Tasks/ Errors										
UE1										
UE2					☑ P3/ P4/ P5					
UE3			☑ P1					☑ P2	☑ P1	
UE4										
UE5										
UE6					☑ P4/ P5/ P6				☑ P1/ P6	
Severity					Low			Low	Low	

*P – Participant Code, T – Task Code, UE – Usability Error Code, Sp – Sprint code

2) *Design feedback*: Table V summarises the experts' feedback based on the remarks and discussion session at the sprints meeting session. The summary is divided into four sections: challenges, improvements, questions to consider, and interesting implemented features. The main areas that need more attention, according to the feedback, are the collection operation, the payment process for collectors, and features for providing feedback on waste quality. As a result, the following actions were put in place.

- Collection operation: using a dynamic QR code, food waste collection at pickups and sending locations could be improved. A dynamic QR code will change over time, making it easier to capture time accurately. As a result, the payment process can be calculated based on the recorded time.
- Payment procedure: the discussion in this section resulted in a manual operation for paying the collectors' commission on waste delivery. The delivery fee will be calculated based on the delivery rate per kilometre and the weight of the waste. The landfill will make weekly payments to collectors. All payment transactions will be recorded; only payment transfers will occur outside the system.
- Providing feedback: it is crucial to rate the collected waste as the land site needs to produce a high quality of decomposed for BSF. Preparation of food waste has to follow the given guideline to avoid receiving food waste that does not meet the standard, and landfills must rate the waste. Any non-compliance should result in negative feedback, and the household or restaurant will be barred from sending waste for an extended period.

The collectors and household/restaurant modules were prioritized following the sprint design meeting. The justification for this was that the functionalities were directly related to the collections processes, which are the application's key features for the key users. After considering all the feedback from the sessions, a series of design concepts were created later to demonstrate the workflow development process.

3) *Implementation of agile UX approach*: Sprint activities are remarkably beneficial in improving the system's designs and functionality when using an Agile UX approach. On the other hand, the concepts can be easily moulded because they are all aimed at the same goal. During the sprints, participants with varying expertise and skill sets contributed a wide variety of ideas and criticism. Nonetheless, at the end of each sprint session, all participants reached an agreement.

TABLE V. FEEDBACK COLLECTED FROM SPRINT ACTIVITIES

Challenges	Viewpoints and Improvements
<ul style="list-style-type: none"> - Getting households to use the apps, generate the QR code, and use the code every time collectors collect waste may be difficult for some users who experience difficulties with mobile technology. - How can landfills managers ensure that households and restaurants provide high-quality Waste for BSF and improved decomposition? - Paying collectors via the system immediately after completing a delivery is more efficient than a manual payment. However, it may be difficult for the landfill manager to control the quality of the waste while also managing a fair collection assignment for all collectors. - Gamification in a mobile application required some time to see the effect on collector motivation and behaviour. 	<ul style="list-style-type: none"> - Reminders and notifications should be used when collectors reach their destination (pick up and send locations). - Payment workflows should be improved. - Gamification features should be explained in detail once all data has been integrated. - Data visualization on the collector's dashboard could assist them in effectively synchronizing their work. - Could external devices such as scanners be incorporated in determining the waste quality? - The app's purpose is to make waste collection and distribution easier. It has made a gig platform job opportunity available to local communities. - A reward is an interesting method to appreciate user involvement - Sprint activities gradually demonstrate the system from concept idea to design step-by-step.
Questions to Ponder	Interesting Features
<ul style="list-style-type: none"> - Local community involvement in catchment teaches good waste management ethics and environmental care surroundings. How do you ensure the information provided and feedback to them is actionable? - Is it possible to shorten the work and process of decomposition by using mobile apps? 	<ul style="list-style-type: none"> - Gamification elements are used to encourage users to participate in collecting and delivering food waste. - The designs are acceptable, and the concepts are considered to make sense as they adequately address the requirements. - Feedback and ratings would be relatively beneficial in nurturing households and restaurants in providing quality food waste.

Generally, communication and collaboration are the main challenges a company faces in Malaysia when using a UX approach [21, 23]. The programmers faced various challenges regarding the technical aspects, particularly those concerning application design and development. The challenges elicited when connecting the system's requirements and designs between the end-users, the software developers, and the software vendor. The developers and user experience practitioners can become frustrated, particularly when openness to sharing the system problems and solutions is limited to the knowledge or skills of developers and practitioners [24]. In such cases, it would expose their inability to solve problems during the development process. Developers might also experience high pressure in providing seamless progress [21]. However, this situation could be because of several independent factors implemented in various ways. For example, a collaboration between developers of different level of skills level to complete a successful system, input from ethnic diversity provides a different insight into the system implementation, and different levels of technology-savvy

among users help to train users to be more competent. Project progress sharing will make the project more visible and transparent. We argue that this variety should provide the team with more valuable input to the projects, mainly in the case of the Malaysian context. These situations have encouraged collaboration, allowing further exploration of the team's creativity. Looking on the positive side, the team could alleviate the challenges with careful team design and well-planned work.

As UX and agile approaches are heavily involved with users in every phase of system development thus, regardless of individual limitations, cooperation from each member is imperative towards achieving the project's aim. The authors anticipated that using an agile UX approach would produce a great system preferred by the users.

V. SYSTEM DESIGN MODEL

The system requirements and designs are reviewed before encapsulating them in a design model specifically for developing such applications for managing local food waste. The model consists of the required components to improve BSF decomposition in the Malaysian context. This model should be used as a guideline when creating similar applications or applying the development approach to software applications.

The model was created to reflect the application's features, purposes, the environment in which it will be used, and its requirements. The applications' core ideas were extracted and summarized from requirements activities to designs and testing into a model. As a result, the model comprises five distinct components, as illustrated in Fig. 7. For each of the components in the model:

- **Motivation:** As the goal of the apps was to have an interactive application that can support food waste management processes and get local communities involved, motivation elements and a rewards system should be used to connect all the components in the model.
- **Stakeholder:** The involvement of stakeholders is critical because they must understand their responsibilities and act accordingly.
- **Systematic and user friendly:** A standard system application should be designed with a user-friendly feature, and the system should be implemented systematically across all aspects of the application. This includes a map showing the collection route and a map that can track collector movement, data visualization to show the waste collection activities and disposal, the payment process, and the rewards system.
- **Job opportunity:** From the application development process, the application should be able to provide users with job opportunities or, at the very least, a side income opportunity. Collectors are also known as riders, and providing a gig job has become a popular trend in the global economy.

- Eco-friendly approach: The applications should promote an environmentally sustainable attitude by raising awareness of safe and proper food waste disposal practices. This strategy will both educate and develop communities indirectly.

The project management team can better plan for system design and development with this model. The proposed system design concept was generally well welcomed. The participants expressed their excitement to see the result of food waste collection operations after the system was fully implemented.

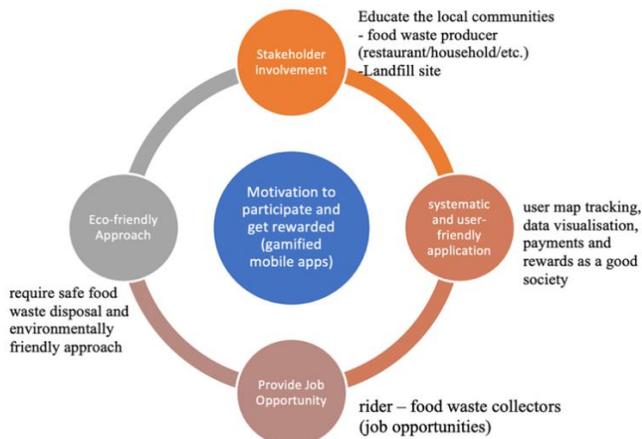


Fig. 7. The Mobile Apps Model for Food Waste.

VI. CONCLUSION

Proper food waste disposal utilizing BSF is one way to support the goal of having and keeping a clean and safe environment. The usage of an assistive tool, such as a gamified mobile application for food waste management, is thought to be necessary to ease the disposal processes. This method would require the involvement of the surrounding communities in order for everyone to contribute to a safe environment. Also, by providing this type of application, communities will be able to learn and adapt to proper and safe food waste disposal in their regular lifestyle. However, waste preparation must meet the stated criteria to maintain decomposition quality. It will also assist local governments in monitoring food waste compliance using the application.

This study describes the process of developing a mobile-based application for managing food waste, focusing on the preparation and collection processes. This study's work was based on the Agile UX methodology, particularly its adoption in the Malaysian context of information system development. In the entire process, two teams are involved: agile and UX. The developed design was subjected to usability testing to ensure that the design and requirements were in sync. Feedback on the process flow and its designs aided in achieving the system's goal. The system model was developed as a guideline when developing a similar application resulting from the design process. Overall, the application designs and features were agreed upon for implementation.

Of course, this study has several limitations. One of them was locating and assigning a time for a sprint meeting to all participants. However, an online meeting appears to be more

convenient for all parties due to the pandemic. It seems to be more effective in moving from one sprint to another. Then, because working from home has become the new norm, various distractions to complete the work sometimes require time extension, dragging the project timeline. However, the authors believe this is a low-risk situation that can be managed until the project is completed. In future work, the researcher will conduct acceptance testing and assess the effectiveness of the application in the actual environment with the actual users. A study on how the application could affect the BSF industry in a bigger picture, i.e., production and business, should also be conducted. At present, the study's findings indicate that the proposed system design and model are feasible and would contribute positively to food waste and the BSF ecosystem.

ACKNOWLEDGMENT

The authors would like to thank the University Malaysia Sabah for the financial support and opportunities in conducting the research, and the participants who participated in completing this study.

REFERENCES

- [1] Abd El-Hack, M. E., Shafi, M. E., Alghamdi, W. Y., Abdelnour, S. A., Shehata, A. M., Noreldin, A. E., Ragni, M. "Black soldier fly (*Hermetia illucens*) meal as a promising feed ingredient for poultry: A comprehensive review". *Agriculture (Switzerland)*, 10(8), pp. 1–31, 2020. <https://doi.org/10.3390/agriculture10080339>.
- [2] Tomberlin, J. K., & van Huis, A. "Black soldier fly from pest to "crown jewel" of the insects as feed industry: An historical perspective". *Journal of Insects as Food and Feed*, 6(1), pp. 1–4, 2020. <https://doi.org/10.3920/JIFF2020.0003>.
- [3] Ramli N., Rosli N.S., Abdul Wahap F., Wan Nawawi W.N., Mohd Abd Majid H.A. "Acceptance of Food Waste Recycling Products among Public toward Sustainable Food Waste Management". In: Kaur N., Ahmad M. (eds) *Charting a Sustainable Future of ASEAN in Business and Social Sciences*. Springer, Singapore, pp. 391–401, 2020. https://doi.org/10.1007/978-981-15-3859-9_33.
- [4] Kim, C.-H.; Ryu, J.; Lee, J.; Ko, K.; Lee, J.-y.; Park, K.Y.; Chung, H. "Use of Black Soldier Fly Larvae for Food Waste Treatment and Energy Production in Asian Countries: A Review". *Processes*, 9, pp. 161, 2021.
- [5] Jain, B., Khosla, A., Chand, K., Ahuja, K. "The Pursuit of Zero Food Waste: A Gamified Approach Promoting Avoidance of Dormitory Mess Food Wastage in Educational Institutions". *Global Initiatives for Waste Reduction and Cutting Food Loss*. IGI Global, pp. 243–267, 2019.
- [6] Ishimura, Y., & Takeuchi, K. (2018). "Where Did Our NIMBY Go? The Spatial Concentration of Waste Landfill Sites in Japan". Available online: <http://www.econ.kobe-u.ac.jp/RePEc/koe/wpaper/2018/1818.pdf> (Accessed on October 2, 2021).
- [7] Watanabe, K., Okayama, T., Siti Khadijah, A. G., Tetsuya, A., Cristina, L., Iderlina, M.-B., Adelia, L. (2018). An exploratory study on waste management in Southeast Asian megacities and Tokyo. *Teikyo Journal of Sociology*, 2018, 31, pp. 115–128.
- [8] The World Bank. (2019). *Understanding Poverty Urban Development: Brief: Solid Waste Management*. Available online: <http://www.worldbank.org/en/topic/urbandevelopment/brief/solid-waste-management> (Accessed on October 3 2021).
- [9] Abd Ghafar S.W. "Food Waste in Malaysia: Trends, Current Practices, and Key Challenges, Policy Articles in Food and Fertilizer Technology Centre for the Asian and Pacific Region". Available online: <https://ap.iftc.org.tw/article/1196> (Accessed November 1, 2021).
- [10] Bernama. (2020, October 20). "SWCorp: Food waste drops during MCO, rises again soon after". *New Straits Times*. Available online: <https://www.nst.com.my/news/nation/2020/10/633738/swcorp-food-waste-drops-during-mco-rises-again-soon-after-Food-wastage-is-a-perpetual-waste-generated-in-Malaysia-daily>. (Accessed November 1, 2021).

- [11] Dalilawati Zainal, Khana Azwar Hassan. "Factors Influencing Household Food Waste Behaviour in Malaysia", *International Journal of Research in Business, Economics and Management*, vol.3 (3), pp. 56-71, 2019.
- [12] Jarjusey, F., Chamhuri, N. "Consumers' Awareness and Knowledge about Food Waste in Selangor Malaysia". *International Journal of Business and Economic (IJBEA)*, vol. 2(2), pp. 91-97, 2017. DOI: 10.24088/IJBEA-2017-22002.
- [13] Varhana, N.A., Faliasthiunus, M.F., Ulfah, P.K. "Traceable Waste System: Konsep Pelacakan Dan Pengelolaan Sampah Menuju Indonesia Bersih Bebas Sampah 2025". *Journal Ilmiah Penalaran dan Penelitian Mahasiswa*, vol. 4, no. 2. pp. 91-100, 2020.
- [14] Gunawardane, M., Pushpakumara, H., Navarathne, E., Lokuliyana, S., Kelaniyage, K., & Gamage, N., (2019). "Zero Food Waste: Food wastage sustaining mobile application", In proceedings of the 2019 International Conference on Advancements in Computing (ICAC), Malabe, Sri Lanka, 5-7 December 2019, pp. 129-132. doi: 10.1109/ICAC49085.2019.9103370.
- [15] Putri Sekar Melati, Sulistinyah Suwaka Putri, Rossy Andini Herindra Putri, Lilit Rusyati. "Android Based Application Fawless (Food Assist Wasteless) As Innovative Solution on Reducing Food Waste". In Proceedings of the 7th Mathematics, Science, and Computer Science Education International Seminar, MSCEIS 2019, Bandung, West Java, Indonesia. 12 October 2019. <http://doi.org/10.4108/eai.12-10-2019.2296470>.
- [16] Koivisto, J., & Hamari, J. "The rise of motivational information systems: A review of gamification research". *International Journal of Information Management*, vol. 45(October 2018), pp. 191–210, 2019. <http://doi.org/10.1016/j.ijinfomgt.2018.10.013>.
- [17] Ulla Santi, Ari Happonen, Harri Auvinen. "Digitalization boosted recycling: Gamification as an inspiration for young adults to do enhanced waste sorting". In proceeding of 13th International Engineering Research Conference, Subang Jaya, Malaysia. 27 November 2019, <https://doi.org/10.1063/5.0001547>.
- [18] Ali Fawzi Najm Al-Shammari, Ahmed Fadhil. "FoodWise: Geolocalised Food Wastes Tracking and Management". *Kerbala Journal for Engineering Science*, Vol. 0 (2), 2020.
- [19] Kieffer, A., Ghouti, A., & Macq, B. "The Agile UX Development Lifecycle: Combining Formative Usability and Agile Methods". Proceedings of the 50th Annual Hawaii International Conference on System Sciences HICSS, Waikoloa, Hawaii, USA, 4 – 7 Jan 2017, pp. 577 – 586.
- [20] Akbar R., Khan A.R., Adnan K. "Software Development Process Evolution in Malaysian Companies". In: Sharma N., Chakrabarti A., Balas V. (eds) *Data Management, Analytics, and Innovation. Advances in Intelligent Systems and Computing*, Springer, Singapore, vol. 1042, 2020. https://doi.org/10.1007/978-981-32-9949-8_10.
- [21] Ma'arif, Muhamad & Shahar, Siti Mariam & Yusof, Mohd Fikri Hafifi & Mohd Satar, Nurhizam. "The Challenges of Implementing Agile Scrum in Information Systems Project", *Journal of Advanced Research in Dynamical and Control Systems*, Vol. 10, 09-Special Issue, pp. 2357-2363, 2018.
- [22] Sutherland, J. Harrison, N. and Riddle, J. "Teams That Finish Early Accelerate Faster: A Pattern Language for High Performing Scrum Teams". In Proceedings of the 47th Annual Hawaii International Conference on System Sciences HICSS, Waikoloa, Hawaii, USA, 6 – 9 Jan 2014, pp. 4722-4728.
- [23] Hussein, I., Hussain, A., Mkpojiogu, E.O.C., & Nathan, S.S. "The state of user experience design practice in Malaysia". *International Journal of Innovative Technology and Exploring Engineering*, vol. 8(8S), pp. 491-497, 2019.
- [24] Hussain, A., Mkpojiogu, E.O.C., & Idyawati, H. "Assessing the frustrations of practicing user experience design (UXD) among the UXD community of practice in Malaysia: A Netnographic Approach", *An International Journal of Advanced Computer Technology, COMPUSOFT*, 8(8), pp. 3347-3355, 2019. DOI: 10.6084/ijact.v8i8.913.

Entanglement Quantification and Classification: A Systematic Literature Review

Amirul Asyraf Zahir¹, Mohd Ilias M Shuhud³

Faculty of Science and Technology
Universiti Sains Islam Malaysia
Negeri Sembilan, Malaysia

Siti Munirah Mohd^{2*}

Genius Insan College
Universiti Sains Islam Malaysia
Negeri Sembilan, Malaysia

Bahari Idrus⁴

Center for Software Technology and Management, Faculty of
Information Science and Technology
Universiti Kebangsaan Malaysia, Selangor, Malaysia

Hishamuddin Zainuddin⁵

Department of Physics, Faculty of Science
Universiti Putra Malaysia
Selangor, Malaysia

Nurhidaya Mohamad Jan⁶

Genius Insan College
Universiti Sains Islam Malaysia
Negeri Sembilan, Malaysia

Mohamed Ridza Wahiddin⁷

Cybersecurity & Systems Unit
Universiti Sains Islam Malaysia
Negeri Sembilan, Malaysia

Abstract—Quantum entanglement is one of the essences of quantum mechanics and quantum information theory. It is a physical phenomenon in which entangled particles remain correlated with each other regardless of the distance between them. Quantum entanglement plays a significant role in areas such as quantum computing, quantum cryptography, and quantum teleportation. Quantifying entanglement is important for determining the depth of the entanglement level and has an impact on quantum information tasks performance. Entanglement classification is critical in quantum information theory for determining the class of states in a quantum system. The entanglement classification of two qubits as separable or entangled has been established. The classification of multiqubit entanglement is more challenging, especially in higher-qubit systems. The goal of this study is to identify different established measurements for entanglement quantification and entanglement classification methods through a systematic literature review. Indexed articles between 2017 and 2021 were selected as secondary resources from several sources based on specific keywords. This study presents a conceptual framework of entanglement quantification and classification based on previous studies.

Keywords—Entanglement quantification; quantum entanglement; entanglement classification; quantum measurement

I. INTRODUCTION

Quantum entanglement is one of the most studied features in quantum mechanics that is critical to quantum information processing [1] in areas such as quantum teleportation, quantum cryptography, and quantum computing. In quantum computing, quantum entanglement plays a vital role in demonstrating the superiority of a quantum computer over its classical counterpart [2]. Although interest in quantum entanglement has grown over the years, knowledge of the

phenomenon is still limited, especially in higher-dimensional systems [3].

Entanglement quantification is a process of determining the level of entanglement and the intactness of a system and characterizing it. It is a fundamental problem in quantum information theory, especially in multipartite settings where the complexity increases with the number of subsystems involved [3, 4]. Some known entanglement quantification measurements are concurrence [5], Schmidt decomposition [6], negativity [7], and entanglement of formation [8].

The most well-known protocols in entanglement classification are local unitary (LU), local operations and classical communication (LOCC), and stochastic local operations and classical communication (SLOCC). The process of entanglement classification is an open problem in quantum information theory. Entanglement classification is established as either separable or entangled in a two-qubit system. The classification becomes more complex as the number of qubits in the system grows. For example, the classification in a three-qubit system is one separable, three biseparable, and two genuinely entangled states (GHZ and W) under SLOCC [9]. The classification for n -qubit ≥ 4 is understudied and it is even more complicated due to the infinite number of classes under SLOCC [10-12].

Entanglement classification is used to categorize the class for complex structures starting from $n = 3$ qubits. It will be complemented with quantification because the measure that has been by quantification determines the degree of entanglement of each state.

This study followed a set of guidelines to identify existing entanglement quantification and classification methods, and propose a framework for the methods. The paper is organized

*Corresponding Author.

as follows. The research methodology is detailed in Section II. Section III covers the results and discussions. Section IV concludes the study.

II. METHODOLOGY

This section discusses the publication standard used in this study, the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA). The following topics are thoroughly discussed: (1) PRISMA review protocol, (2) research question formulation, (3) systematic searching strategy, (4) quality appraisal, and (5) data extraction and analysis.

A. PRISMA Review Protocol

PRISMA [13] was used as the review protocol in this systematic literature review. This review protocol served as a guideline for conducting a systematic literature review by formulating research questions, identifying the inclusion and exclusion criteria in a systematic searching strategy, conducting a quality appraisal of selected articles, and critical data extraction and analysis over a specified period. The scope of this study is entanglement quantification and classification.

B. Research Question Formulation

The systematic literature review is guided by the research questions developed during the preliminary phase. The following research questions were formulated in accordance with the research objective of presenting a conceptual framework of entanglement quantification and classification: (1) What are the established methods of entanglement quantification? (2) What are the established methods for classifying quantum entanglement? (3) What is the preferred alternative method of entanglement quantification and classification?

C. Systematic Searching Strategy

The systematic searching strategy of this study consists of three steps: identification, screening, and eligibility.

1) *Identification*: Identification is an important process as it determines which articles are relevant to the review. The articles for the study were primarily drawn from the databases of two powerful multidisciplinary search engines, Scopus and Web of Science (WOS), as well as an additional database, Google Scholar. A comprehensive search was conducted using the field tags “TITLE-ABS-Key” (title, abstract, and keywords) in Scopus and “TS” (topic) in WOS with the keywords entanglement quantification, quantum entanglement, entanglement classification, and quantum measurement.

The search strings were created to search for related articles in both databases. The searches were conducted from November to December 2021 (see Table I). A manual search was conducted using Google Scholar, with handpicked related articles derived from the same keywords as in Scopus and WOS. Based on the systematic searching of Scopus and WOS, a total of 13,659 potential related articles were identified and 1,934 articles were downloaded. Additionally, 8 articles were selected from Google Scholar for further analysis. The search results from Scopus and WOS are displayed in Fig. 1.

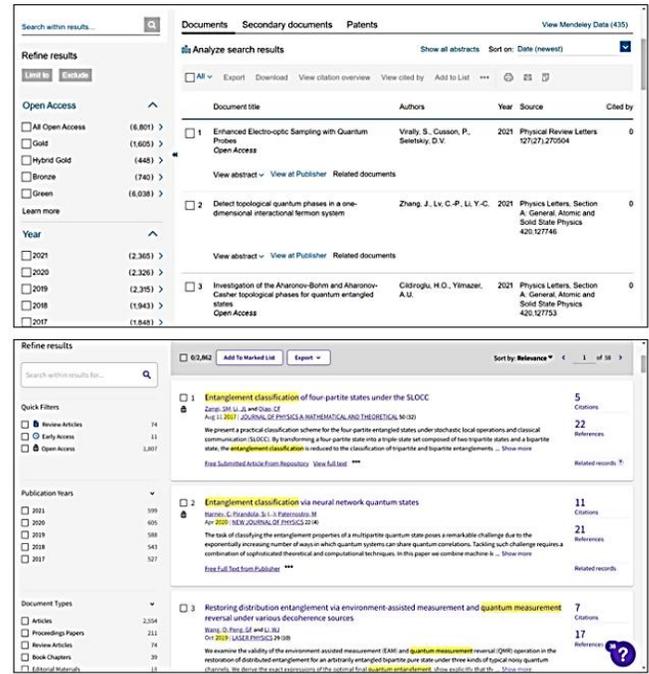


Fig. 1. Search Results in Scopus and WOS.

2) *Screening*: This systematic literature review examined indexed articles on entanglement quantification and classification published between 2017 and 2021. The five-year period was chosen because of the maturity of the subject [14]. 1,847 of the total 1,942 downloaded articles were excluded due to duplication by title review and abstract review. During the screening stage, the remaining 95 articles were validated to ensure they met the inclusion and exclusion criteria. The inclusion and exclusion criteria are the subject matter, literature type, language of the article, and year of publication (see Table II). Articles that are unrelated to entanglement quantification and classification were excluded from this study. As a precaution against misunderstanding and mistranslation, only articles written in English were considered. After this stage, 62 articles met the inclusion and exclusion criteria.

TABLE I. SYSTEMATIC REVIEW PROCESS SEARCH STRING

Database	Search string
Scopus	TITLE-ABS-KEY (“entanglement quantification” OR “quantum entanglement” OR “entanglement classification” OR “quantum measurement”)
WOS	TS= (“entanglement quantification” OR “quantum entanglement” OR “entanglement classification” OR “quantum measurement”)

TABLE II. INCLUSION AND EXCLUSION CRITERIA

Inclusion	Exclusion
Articles on the subject matter of “entanglement quantification” and “entanglement classification”	Articles written in languages other than English
Indexed journal articles	Articles published before 2017

3) *Eligibility*: In the third stage of the systematic searching strategy, the remaining 62 articles from the screening stage were reviewed again for suitability for this study. After a thorough examination, 27 articles were removed since their research direction or theme did not focus on entanglement quantification and classification. The remaining 35 articles were then prepared for a quality appraisal (see Fig. 3).

D. Quality Appraisal

The 35 selected articles were sent to an expert in the field for quality appraisal to ensure that only high-quality articles were used in the review. According to [15], the remaining articles should be ranked as high, moderate, or low quality, with only high and moderate-quality articles being included in the review. To meet the quality standard, the expert concentrated on specific elements such as the theme, objective, and results of the articles. Following the appraisal, the expert determined that all 35 articles were suitable for the review.

E. Data Extraction and Analysis

In-depth analysis was used to extract relevant data from the articles by first analyzing the abstract, then the discussion and conclusion, and finally the body for any other relevant

information. The extracted data were tabulated in Microsoft Word. The articles were divided into five categories based on the year they were published (see Fig. 2). There are 8 articles published in 2017, 8 articles in 2018, 3 articles in 2019, 5 articles in 2020, and 11 articles in 2021.

The main themes in the articles from the extracted data are entanglement quantification and entanglement classification. In the following section, we will go over a few of the methods that were discovered.

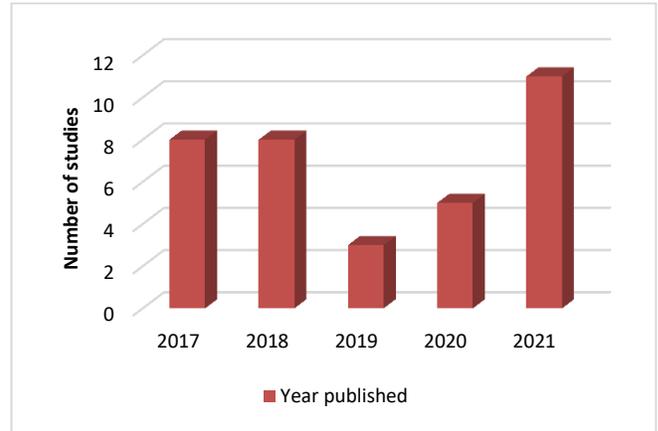


Fig. 2. Articles Group by Year Published.

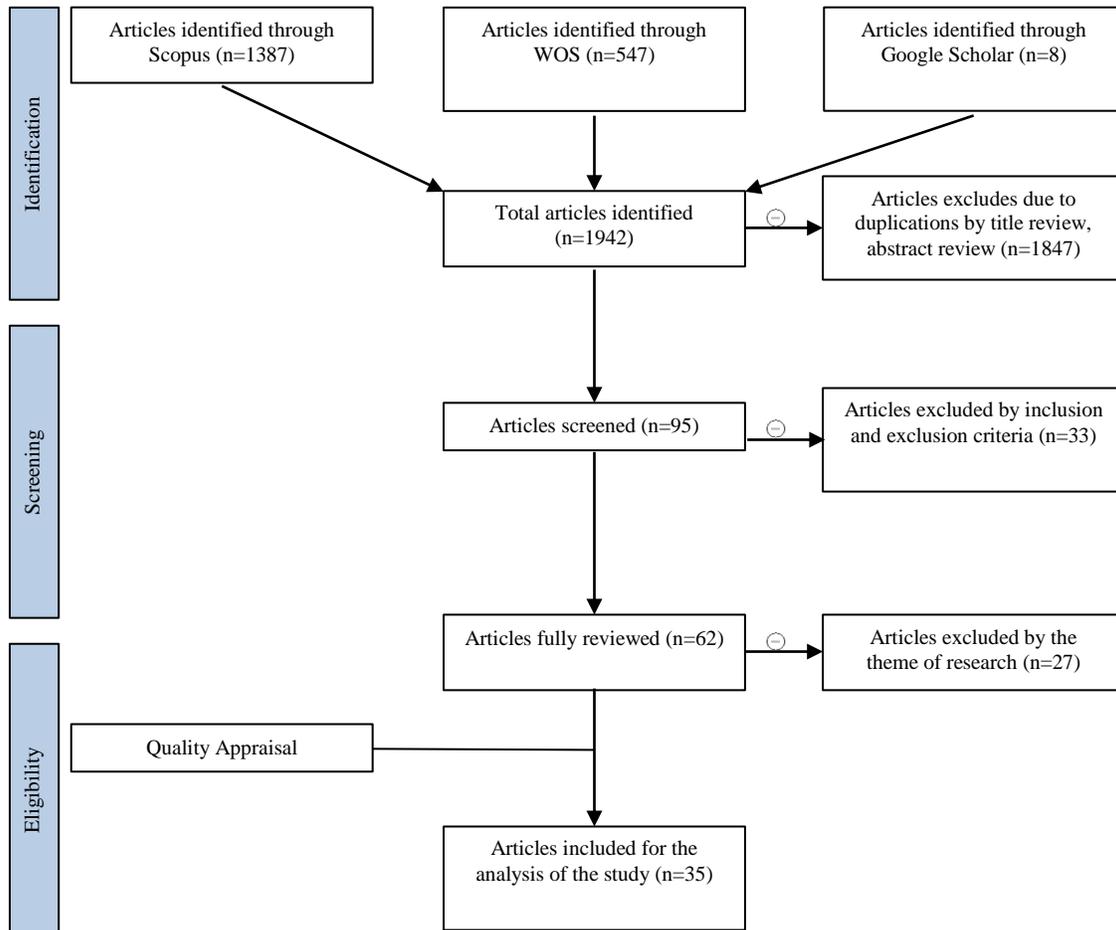


Fig. 3. Article Selection Process.

III. RESULT

This section discusses the identified themes, entanglement quantification, and entanglement classification in previous studies (see Table III and Table IV). A proposed conceptual framework of entanglement quantification and classification has been developed and is presented as a reference for future work (see Fig. 6).

A. Entanglement Quantification and Entanglement Classification Methods

There are 19 entanglement quantification methods established from previous studies, as shown in Table III. Some of these methods are compounded (additional variables or adaptation) such as base or compounded concurrence [4, 5, 8, 16-19], base or compounded negativity [4, 7, 8, 19], base or compounded entanglement of formation [4, 5, 8, 20], base or compounded convex-roof measures [6, 8, 21], base or compounded tangle [4, 5, 17], base or compounded entanglement witness [22, 23], relative entropy of entanglement [20, 21], and Schmidt decomposition [6, 20]. Other established entanglement quantification methods are: (1) Tsallis-q entanglement measure, (2) k-entanglement measure, (3) entanglement of assistance, (4) supervised machine learning, (5) linear entropy of entanglement, (6) an extension of an entanglement measure for the mixed state from the measure of a pure state, (7) global nonselective projective measurement, (8) Gramian operators, (9) exact PPT entanglement cost, (10) operational entanglement monotone approach, and (11) entanglement of formation.

These methods were grouped into the following clusters:
Cluster 1: Concurrence; Cluster 2: negativity; Cluster 3: Entanglement of formation; Cluster 4: Convex-roof measures; Cluster 5: Tangle; Cluster 6: Entanglement witness; Cluster 7: Relative entropy of entanglement; and Cluster 8: Schmidt decomposition. Fig. 4 depicts the clusters of entanglement quantification methods. Concurrence is the most commonly used entanglement quantification method, followed by negativity. Base or compounded concurrence and negativity are widely regarded as simple and direct measures of entanglement compared to other listed methods in this research.

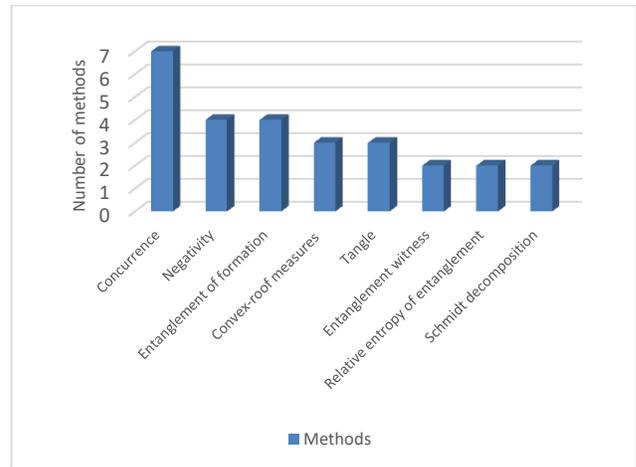


Fig. 4. Entanglement Quantification Methods Cluster.

TABLE III. ENTANGLEMENT QUANTIFICATION METHODS FROM PREVIOUS STUDIES

Source	Methods	Qubit system		Quantum state			Remarks
		BP	MP	PU	MX	AR	
[5]	Concurrence, tangle, Tsallis-q entanglement measure, entanglement of formation, squared concurrence	⊗			⊗		N-qubit = 2 High dimensional system
[22]	Witness operator	⊗			⊗		N-qubit = 2 High dimensional system
[6]	Ent (PU – Schmid decomposition; MX – Convex roof extension)	⊗	⊗	⊗	⊗		N-qubit = 2, 3, n MX - Not applicable for n-partite and above
[4]	Negativity tangle, entanglement of formation tangle, concurrence tangle		⊗	⊗	⊗		N-qubit = 3, 4, n
[16]	Ent-concurrence	⊗	⊗	⊗	⊗		N-qubit = 2, 3, 4 Detects entanglement in reduced and full states
[24]	Entanglement of assistance		⊗	⊗			N-qubit = 3
[3]	Supervised machine learning		⊗	⊗			N qubit ≤ 8
[17]	A family of multipartite entanglement - concentratable entanglements (n-tangle, concurrence, linear entropy of entanglement)		⊗	⊗			N-qubit = 3
[25]	An extension of an entanglement measure for the mixed state from the measure of pure state	⊗		⊗	⊗		N-qubit = 2
[7]	Negativity/global nonselective projective measurement	⊗		⊗	⊗		N-qubit = 2
[20]	Fidelity based distance (relative entropy, entanglement of formation, and Schmidt decomposition)	⊗	⊗	⊗	⊗		N-qubit = 2, 3

[26]	Ent Detector – computational toolbox (Gramian operators)	✳		✳	✳		N-qubit = 2
[23]	Quantitative measurement-device-independent entanglement witness (MDI-EW)	✳				✳	N-qubit = 2 Extendable to multipartite (n-qubit)
[27]	Operational entanglement monotone approach	✳		✳	✳		N-qubit = 2
[21]	Axiomatic approach - Convex roof entanglement measures (the relative entropy of entanglement, the negativity, the logarithmic negativity, and the logarithmic convex-roof extended negativity)	✳		✳	✳		N-qubit = 2
[18]	Concurrence		✳	✳			N-qubit = 7
[19]	Concurrence of assistance and negativity of assistance	✳	✳	✳			N-qubit = 2, 3, n
[8]	Concurrence, negativity, convex-roof extended negativity, entanglement of formation	✳	✳	✳	✳		N-qubit = 2, 3
[28]	k-entanglement measure and exact PPT entanglement cost	✳	✳	✳		✳	N-qubit = 2, 3

BP = Bipartite, MP = Multipartite, PU = Pure State, MX = Mixed State, AR = Arbitrary State

TABLE IV. ENTANGLEMENT CLASSIFICATION METHODS FROM PREVIOUS STUDIES

Source	Methods	Protocol	Qubit system		Quantum state			Remarks
			BP	MP	PU	MX	AR	
[3]	Supervised machine learning	LOCC		✳	✳			N-qubit ≤ 8
[17]	A family of multipartite entanglement - concentratable entanglements (n-tangle, concurrence, linear entropy of entanglement)	LOCC		✳	✳			N-qubit = 3
[25]	An extension of an entanglement measure for the mixed state from the measure of the pure state	LOCC	✳		✳	✳		N-qubit = 2
[7]	Negativity / Global nonselective projective measurement	LOCC	✳		✳	✳		N-qubit = 2
[26]	Ent Detector – computational toolbox (Gramian operators)	LU	✳		✳	✳		N-qubit = 2
[23]	Quantitative measurement-device-independent entanglement witness (MDI-EW)	LOCC	✳				✳	N-qubit = 2
[27]	Operational entanglement monotone approach	LOCC	✳		✳	✳		N-qubit = 2
[21]	Axiomatic approach - Convex roof entanglement measures (the relative entropy of entanglement, the negativity, the logarithmic negativity, and the logarithmic convex-roof extended negativity)	LU	✳		✳	✳		N-qubit = 2
[28]	K-entanglement measure and exact PPT entanglement cost	LOCC	✳	✳	✳		✳	N-qubit = 2, 3
[29]	A set of operators (contains only Pauli matrices)	SLOCC		✳	✳	✳		N-qubit = 3
[9]	Witness operator	SLOCC		✳	✳			N-qubit = 3
[10]	Algebraic geometry (SLOCC invariants – secant varieties) – k-secants and ℓ-multiranks	SLOCC		✳	✳			N-qubit = 5
[30]	Entanglement polytope	LU		✳	✳			N-qubit = 3
[31]	Maximal Schmidt rank	SLOCC		✳	✳			N-qubit = 3
[11]	Singular value decomposition	SLOCC		✳	✳			N-qubit = 4
[32]	Grover’s algorithm, Shor’s algorithm, Quantum Fourier Transform	SLOCC	✳	✳	✳			N-qubit = 2, 3, 4
[33]	Inductive classification approach	SLOCC		✳	✳			N-qubit = 4
[34]	Invoking the proportional relationships for spectrums and standard Jordan normal forms	SLOCC	✳	✳	✳			N-qubit = 2, 3, 4
[35]	Pauli z-operators	SLOCC		✳	✳			N-qubit = 3
[36]	Bell inequalities	LU		✳	✳			N-qubit = 3
[37]	Special unitary group	LU		✳	✳			N-qubit = 3
[38]	Algebraic geometry	SLOCC	✳	✳	✳			N-qubit = 2, 3, 4
[39]	Separable neural network quantum state	LOCC		✳		✳		N-qubit = 3
[40]	Integer partitions	SLOCC		✳	✳			N-qubit = 4
[41]	Single polynomial entanglement measure	SLOCC		✳	✳			N-qubit = 4

BP = Bipartite, MP = Multipartite, PU = Pure State, MX = Mixed State, AR = Arbitrary State, LU = Local unitary, LOCC = Local operations and classical communication, SLOCC = Stochastic local operations and classical communication

Table IV lists 25 published entanglement classification methods. The protocols used in the studies were emphasized instead of the methods. The LU, LOCC, and SLOCC protocols were identified in the articles. As shown in Fig. 5, it was found that SLOCC is the most utilized protocol. This may be due to the fluidity of the SLOCC protocols in classifying entanglement. The protocols were classified into three clusters, as depicted in Fig. 5.

B. Conceptual framework of entanglement quantification and classification

Methods for quantifying and classifying entanglement established in previous studies were thoroughly examined to comprehend the essence of both concepts. The purpose of this research is to develop a conceptual framework of entanglement quantification and classification in bipartite and multipartite systems. The framework was developed following the specifications established in previous studies on the

quantum qubit system and state for entanglement quantification, as well as entanglement classification protocols.

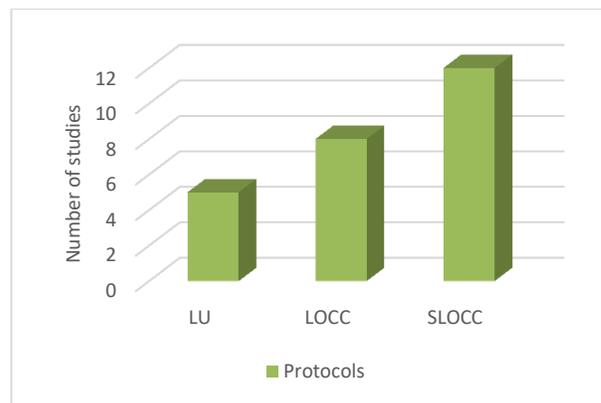


Fig. 5. Entanglement Classification Protocols Clusters.

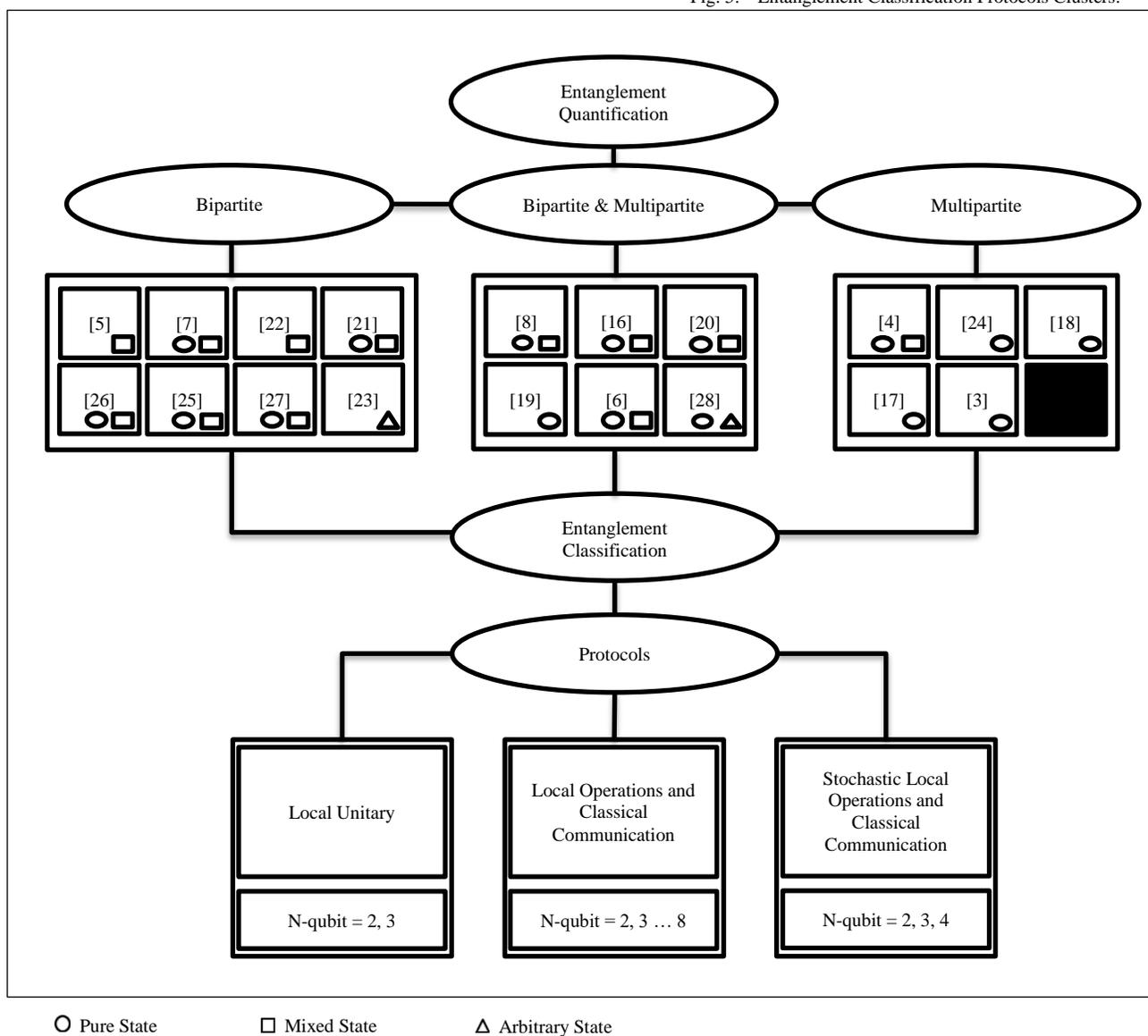


Fig. 6. Proposed Conceptual Framework of Entanglement Quantification and Classification.

IV. CONCLUSION

Even though significant progress has been achieved, entanglement quantification and classification remains a challenging and open problem in quantum information processing, especially in mixed quantum state and when there are many particles (qubits) involved, i.e., n -qubit ≥ 4 .

This study presents several established methods for quantifying and classifying entanglement that have been identified in previous studies. In addition, a conceptual framework of entanglement quantification and classification in bipartite and multipartite systems was developed and presented as a guidance or reference for future work based on one's specific requirements, namely measurement methods, qubit system, quantum state and protocols.

The understanding of the entanglement measures and classification is still considered insufficient. Therefore, further study on entanglement quantification and classification methods based on the proposed conceptual framework is needed to produce a universal quantification measurement and precise classes or families for classification in a higher-qubit and higher-dimensional system.

ACKNOWLEDGMENT

This research is part of a research project supported by the Malaysian Ministry of Higher Education Fundamental Research Grant Nos. FRGS/1/2021/ICT04/USIM/01/1 (USIM/FRGS/KGI/KPT/50121).

REFERENCES

- [1] K. Zhijin et al., Detection and quantification of entanglement with measurement-device-independent and universal entanglement witness, *Chinese Physics B*, 2020. 29.
- [2] A. A. Zhahir et al., Quantum Computing and Its Application, *International Journal of Advanced Research in Technology and Innovation*, 2022. 4 (1): pp. 55-65.
- [3] C. Chen, C. Ren, H. Lin, and H. Lu, Entanglement structure detection via machine learning, *Quantum Science and Technology*, 2021. 6.
- [4] S. P. J. Geetha, K. S. Mallesh, Comparative analysis of entanglement measures based on monogamy inequality, *Chin. Phys. B*, 2017. 26 (5): pp. 50301-050301.
- [5] M. Moslehi, H. R. Baghshahi, and S. Y. Mirafzali, Upper and lower bounds for Tsallis-q entanglement measure, *Quantum Information Processing*, 2020. 19 (11): p. 413.
- [6] S. R. Hedemann, Ent: A Multipartite Entanglement Measure, and Parameterization of Entangled States, arXiv preprint arXiv:1611.03882, 2018.
- [7] M. Zuppardo, R. Ganardi, M. Miller, S. Bandyopadhyay, and T. Paterek, Entanglement gain in measurements with unknown results, 2018.
- [8] X.-N. Zhu and S.-M. Fei, Monogamy properties of qubit systems, *Quantum Information Processing*, 2018. 18 (1): p. 23.
- [9] A. Kumari and S. Adhikari, Classification witness operator for the classification of different subclasses of three-qubit GHZ class, *Quantum Information Processing*, 2021. 20 (9): p. 316.
- [10] M. Gharahi, S. Mancini, and G. Ottaviani, Fine-structure classification of multiqubit entanglement by algebraic geometry, *Physical Review Research*, 2020. 2 (4): p. 043003.
- [11] S. M. Zangi, J.-L. Li, and C.-F. Qiao, Entanglement classification of four-partite states under the SLOCC, *Journal of Physics A: Mathematical and Theoretical*, 2017. 50 (32): p. 325301.
- [12] M. Walter, D. Gross, and J. Eisert, Multi-partite entanglement, 2017.
- [13] M. J. Page et al., The PRISMA 2020 statement: an updated guideline for reporting systematic reviews, *BMJ*, 2021. 372 p. n 71.

- [14] S. Kraus, M. Breier, and S. Dasí-Rodríguez, The art of crafting a systematic literature review in entrepreneurship research, *International Entrepreneurship and Management Journal*, 2020. 16 (3): pp. 1023-1042.
- [15] M. Petticrew and H. Roberts, *Systematic Reviews in the Social Sciences: A Practical Guide*, 2006.
- [16] S. R. Hedemann, Candidates for universal measures of multipartite entanglement, arXiv preprint arXiv:1701.03782, 2017.
- [17] J. L. Beckey, N. Gigena, P. J. Coles, and M. Cerezo, Computable and Operationally Meaningful Multipartite Entanglement Measures, *Physical Review Letters*, 2021. 127 (14): p. 140501.
- [18] S.-K. Chu, C.-T. Ma, R.-X. Miao, and C.-H. Wu, Maximally entangled state and Bell's inequality in qubits, *Annals of Physics*, 2018. 395 pp. 183-195.
- [19] Z.-X. Jin and S.-M. Fei, Tighter monogamy relations of quantum entanglement for multiqubit W-class states, *Quantum Information Processing*, 2017. 17 (1): p. 2.
- [20] Y. Guo, L. Zhang, and H. Yuan, Entanglement measures induced by fidelity-based distances, *Quantum Information Processing*, 2020. 19 (9): p. 282.
- [21] X. Qi, T. Gao, and F. Yan, The verification of a requirement of entanglement measures, *Quantum Information Processing*, 2021. 20 (4): p. 133.
- [22] S. Aggarwal and S. Adhikari, Witness operator provides better estimate of the lower bound of concurrence of bipartite bound entangled states in $d_1 \otimes d_2$ -dimensional system, *Quantum Information Processing*, 2021. 20 (3): p. 83.
- [23] D. Rosset, A. Martin, E. Verbanis, C. C. W. Lim, and R. Thew, Practical measurement-device-independent entanglement quantification, *Physical Review A*, 2018. 98 (5): p. 052332.
- [24] K. Pollock, G. Wang, and E. Chitambar, Entanglement of assistance in three-qubit systems, *Physical Review A*, 2021. 103 (3): p. 032428.
- [25] X. Shi and L. Chen, An Extension of Entanglement Measures for Pure States, *Annalen der Physik*, 2021. 533.
- [26] R. Gierlerak, M. Sawerwain, J. Wiśniewska, and M. Wróblewski, EntDetector: Entanglement Detecting Toolbox for Bipartite Quantum States, 2021, pp. 113-126.
- [27] D.-H. Yu and C.-S. Yu, Quantifying entanglement in terms of an operational way, *Chin. Phys. B*, 2021. 30 (2): pp. 20302-0.
- [28] X. Wang and M. M. Wilde, Cost of Quantum Entanglement Simplified, *Physical Review Letters*, 2020. 125 (4): p. 040502.
- [29] C. Datta, S. Adhikari, A. Das, and P. Agrawal, Distinguishing different classes of entanglement of three-qubit pure states, *The European Physical Journal D*, 2018. 72 (9): p. 157.
- [30] S. Luna-Hernández, Some Remarks on the Local Unitary Classification of Three-Qubit Pure States, *Journal of Physics: Conference Series*, 2020. 1540 (1): p. 012025.
- [31] Y. Li, Y. Qiao, X. Wang, and R. Duan, Tripartite-to-Bipartite Entanglement Transformation by Stochastic Local Operations and Classical Communication and the Structure of Matrix Spaces, *Communications in Mathematical Physics*, 2018. 358 (2): pp. 791-814.
- [32] H. Jaffali and F. Holweck, Quantum entanglement involved in Grover's and Shor's algorithms: the four-qubit case, *Quantum Information Processing*, 2019. 18 (5): p. 133.
- [33] M. Backens, Number of superclasses of four-qubit entangled states under the inductive entanglement classification, *Physical Review A*, 2017. 95 (2): p. 022329.
- [34] D. Li, SLOCC classification of n qubits invoking the proportional relationships for spectrums and standard Jordan normal forms, *Quantum Information Processing*, 2017. 17 (1): p. 1.
- [35] A. Singh, H. Singh, and K. Dorai, Experimental classification of entanglement in arbitrary three-qubit pure states on an NMR quantum information processor, *Physical Review A*, 2018. 98 (3): p. 032301.
- [36] A. Das, C. Datta, and P. Agrawal, New Bell inequalities for three-qubit pure states, *Physics Letters A*, 2017. 381 (47): pp. 3928-3933.
- [37] S. M. Mohd, B. Idrus, H. Zainuddin, and M. Mukhtar, Entanglement classification for a three-qubit system using special unitary groups, SU

- (2) and SU (4), *International Journal of Advanced Computer Science and Applications*, 2019.
- [38] M. Sanz, D. Braak, E. Solano, and I. L. Egusquiza, Entanglement Classification with Algebraic Geometry, *Journal of Physics A Mathematical and Theoretical*, 2017. 50 p. 195303.
- [39] C. Harney, M. Paternostro, and S. Pirandola, Mixed state entanglement classification using artificial neural networks, *New Journal of Physics*, 2021. 23 (6): p. 063033.
- [40] D. Li, Entanglement classification via integer partitions, *Quantum Information Processing*, 2019. 19 (1): p. 27.
- [41] A. Burchardt, G. M. Quinta, and R. André, Entanglement Classification via Single Entanglement Measure, arXiv preprint arXiv:2106.00850, 2021.

Effective Cross Synthesized Methodology for Movie Recommendation with Emotion Analysis through Ranking Score

R Lavanya¹, Dr. B Bharathi²

Department of Computer Science and Engineering, Sathyabama Institute of Science and Technology, Chennai, India^{1,2}
Department of Computing Technologies, School of Computing, SRM Institute of Science and Technology, Chennai, India¹

Abstract—Providing accurate movie recommendations to a user with limited computing capability is a challenging task. A hybrid system offers a good trade-off between the accuracy and computations needed for such recommendations. Collaborative Filtering and Content-Based Filtering are two of the most widely employed methods of computing such recommendations. In this work, a high-efficient hybrid recommendation algorithm is proposed, which deeds users' contour attributes to screen them into various groups and recommends movie to a user based on rating given by other similar users. Compared to traditional clustering-based CF recommendation schemes, our technique can effectively decrease the time complexity, whereas attaining remarkable recommendation output. This approach mitigates the shortcomings of the individual methods, while maintaining the advantages. This allows the system to be highly reactive to new viewer inputs without sacrificing on the quality of the recommendations themselves. Building on other hybrids of a similar kind, our proposed system aims to reduce the complexity and features needed for calculation while maintaining good accuracy and further enhanced by utilizing Sentiment Analysis to rank the movies and take user reviews into consideration, which traditional hybrids do not take into account. Then analysis was performed on the data set and the results show that the proposed recommendation system outperforms other traditional approaches.

Keywords—Recommendation systems; collaborative filtering; styling; content based filtering; implicit feedback; hybrid recommendation; sentiment analysis; singular value decomposition

I. INTRODUCTION

Watching movies is one of the most popular forms of media entertainment. Viewers are incredibly engrossed and invested in the culture of motion films. Recent advancements in technology have enabled the widespread streaming of movies on demand [1]. This, in turn, has added to the popularity and ease of access to movies. There are thousands of movies for one to choose from. These movies are not only segregated by their genre, cast, production teams, direction, and numerous other factors. This makes it particularly difficult to pick a single movie to watch first. Everyone's preference of movies is also subjective and one may not enjoy a movie that another person loves. This creates an ambiguity as it is complex to determine what features the most impactful when are looking for a new movie to watch [2]. Movie recommender systems help combat this by providing recommendations based

on what the user may have already watched by suggesting similar movies. It attempts to figure out what movie a viewer is likely to rate among the highest.

Recommending a movie is not a straightforward task. This is particularly challenging also as the preferences of viewers may be very different. This leads to there being a distribution of niches that are not uniform to be immediately apparent [3]. Consequently, a movie that is not conventionally popular may be preferable to some viewer simply based on their subjective view towards movies in general. This can be tackled effectively by taking into account a large variety of movies and a large amount of them so as to encompass the likes and dislikes of users of all categories [4]. A robust recommender must be able to recommend movies that are more relevant to the user themselves as shown in Fig. 1.

Many techniques have been used to make recommenders effective in this regard and perform well with large data [5]. Recommender systems are a set of algorithms aimed at emulating information processing systems where the end goal is to suggest relevant items to users, items being movies that users watch. The various classification of recommender system is given in Fig. 2. Content Based methods also offer a way to deal with the issue of limited rating data [6]. Content Based methods work by taking into consideration the similarity between the movies themselves. This similarity could be between various aspects of the movies such as the genre, cast and other related data [7].

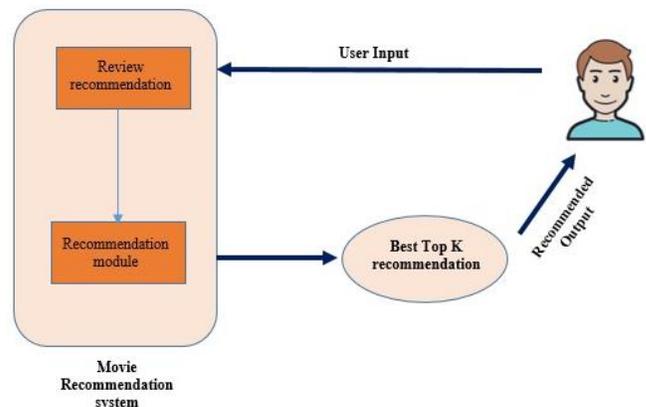


Fig. 1. Schematic of Recommendation System.

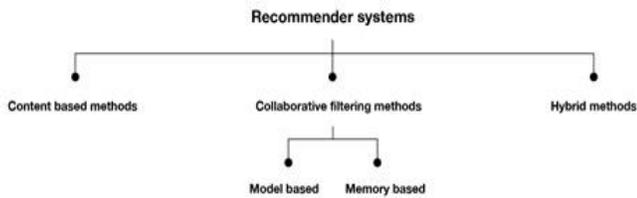


Fig. 2. General Classification of Recommendation System.

A problem that arises with the content-based system is an uncertainty if the system can take the user's behavior towards one genre or source and apply that preference shown to all the other content present in the system. If the engine simply keeps suggesting content similar to what the user is already watching, its value goes down quite a bit in some sectors of recommendation. The Collaborative Filtering approach recommends items by creating a profile for users [8]. The similarity between the profiles is based on the movies rated. If the user rates certain movies high then the algorithm would look for other users who have also rated the same movie highly or in the range of the rating by the user to whom it is to give recommendations. It then predicts what rating the user would give to a movie that they have not seen based on how they have rated the movies they have already seen. Therefore, Collaborative Filtering works well where there is history available for the user ratings on other movies [9]. This however is also a drawback as new users do not have many ratings and this poses a problem for Collaborative Filtering [10]. Based on the advantages, a hybrid system seems to be the most promising approach to mitigate the drawbacks of these common systems and to bring forth their advantages [11]. There are many ways in which Hybrid models can be used: by extrapolating separate content-based recommendation list and collaborative-based recommendation list of predictions and unifying them as one single list; using collaborative-based methods as a primary approach and enhancing them by adding content-based capabilities; using content-based methods as a primary approach and enhancing them by adding collaborative-based capabilities; combining the capabilities and features of both models and creating one single model. This paper is structured as follows: Section 1 emphasis the introduction. Section 2 emphasis the related study and the narrative of the objective. The procedures and resources are discoursed in Section 3 and their exploration and clarifications are exposed in Section 4. The conclusion was clarified in Section 5.

The rest of the paper flows as literature survey in Section II, Section III explain the methodology and implementation process of recommendation system, Section IV describes the comparison and analysis of various recommendation techniques, and Sections V and VI briefs the conclusion and future work.

II. RELATED WORK

Collaborative Filtering algorithm generates a user profile based on the ratings the viewer has given to other movies. If two users have rated similar kind of movies highly, this means that there is a good chance they may prefer similar movies. If the other viewer has already rated other movies, it allows the system to predict if the target viewer will enjoy it or not. This

is the basic premise of how a Collaborative Filtering system would recommend a movie [12]. Collaborative Filtering has been widely employed in many ways, and a lot of academic work has revolved around combining other techniques to boost the performance of such a method. One of the main issues when it comes to Collaborative Filtering is that the computations needed are heavy. Thus in most cases, it has trouble when trying to scale the data up [5]. The method uses Self-Organizing Map Neural Networks [13] to carry out Collaborative Filtering. This method offers a good alternative as the Self Organizing Map Neural Network is not very computationally heavy. While this method worked well for the data, the data itself only comprised of a few dozen movies and about a hundred and seventy users. This makes it hard to judge if the effectiveness of this technique will remain high when faced with a larger dataset, or one with more features.

This multilayer perceptron neural network system works by utilizing the reduction in error in prediction by subjecting the training data to go through multiple passes of a neural network [14]. This method is promising as it does not need to have a deep network for classification. This is particularly effective for polar sentiment data, which will be the focus of the proposed system. This is because for binary classification, a very deep neural network may in fact introduce overfitting to the data. Overfitting occurs when the data trains too well for the test set. This may result in the model performing really well for the trained data but not for the actual target or testing data itself. A shallow network also allows for faster inference time. A personalized Recommendation approach [15] grounded on Three Social Influences, Personal interest means user-item relationship and interpersonal influence and interpersonal interest similarity means user-user relationship of social networks. Probabilistic matrix Factorization makes experiments on the datasets, namely, MovieLens and yelp [16]. Tactically this removes the tricky cold start and data sparsity.

A recommendation system for real estate websites [17] is that it helps consumers in acquiring new properties or homes. Recommendation system is proven by merging case based reasoning (CBR) and Ontology. Former systems supports single characteristic exploration systems but this system support multivalued search system. Sentiment Analysis can also be used along with Collaborative Filtering for better and more inclusive results. The system [18] was trained on data where all the users had given a large number of ratings. This brings into question how well the system would perform where the ratings are limited. Also, the similarities between other features of the movies, such as genre, were not considered in the study.

The author in [19] used a Diverse Collaborative Prediction to combine Collaborative Filtering with Content-Based filtering. This system gave better results than just the individual techniques did, however this method does not consider reviews either. The author in [20] employed an Item-Based Collaborative Filtering model with a Content Based one. The predictions are reached by the TF-IDF method with the nearest neighbor predictions. The MovieLens and Film Trust datasets were used for training and testing this system. The author in [21] used a cosine similarity matrix that showed better

accuracy than that of the other systems, but it also did not take written reviews into consideration.

Another study [22] showed that the Singular Value Decomposition worked well for recommendations. These studies support that the TF-IDF and SVD methods take a lot of heavy computation, but they give some of the most accurate results for recommendations. The author in [23] focused on the hybrid recommendation model which encouraged the user's social data, reviews, and ratings available.

This model of recommendation consists of six processes, review transformation, the feature generation, community prediction, model training, feature blending, and prediction and the last one, evaluation criteria for ontology based recommendations. This [24,27] mainly focuses on a system which helps to provide details analysis about the items which are arranged by the wishes of the similar users. The recommendation system with the proper recommendation for this research will be used in suggesting the item selection system by making a recommendation system with the help of an item-based collaborative filtering methodology. Based on the literature, the associated research challenges are observed.

- Data sparsity may happen due to user/rating matrix is sparse and it is hard to find the users who have rated the same item.
- The existing recommendation technique requires enormous processing time and mostly user is prohibited in getting accurate recommendations that are similar to their profile.

Subsequently it is fortified about the essential for the proposed research to enhance the movie recommendation process competently. The associated objectives are proposed in this research work so as to address few issues in recommendation technique. The contribution of the research comprises.

- To provide enhanced movie recommendation system for the users through an improved hybrid recommendation algorithm combination of user-based CF (UBCF) and item-based CF(BCF) in the context of SVD dimension reduction to improve the speed and quality of recommendation.
- To providing content related to the collection of relevant and irrelevant items for users of online service providers and to recommend movies to users based on user / item base movie ratings.

To enhance the recommendation accuracy in hybrid recommendation system through optimized sentiment analysis for providing more diverse recommendations by satisfying the requirements recommendation features.

III. METHODOLOGY

A. User-Based Collaborative Filtering

User-Based Collaborative Filtering is a technique for predicting which products a user would enjoy based on the ratings provided to that item by other users who share the target user's tastes [1]. Collaborative filtering is used by many

websites to develop their recommendation systems. Steps for Collaborative Filtering with Users: Step 1: Identifying users who are similar to the target user U. The algorithm may be used to calculate similarity between any two users 'a' and 'b'. Step 2: Estimate of an item's missing rating done as follows: Now, the target user may be quite similar to certain people while being very different from others [13, 26]. The proposed system employs a combination of Collaborative Filtering and Content Based Recommendations, further enhanced using Sentiment Analysis to rank the movies.

The movielens dataset is hired in our research paper and collected from the GroupLens [25,28], which contains 20 million ratings for around 27000 different movie titles and has a user ID, movie ID, rating, and timestamp. The Characterization of the movie's content information includes over 54058 records and includes movie ID, title, genre, director, actor, and more. The graph in Fig. 3 represents the relation between categories and the movies rated accordingly.

The data contains a huge amount of reviews. This helps retain most of the movies while reducing the number of users by about a third and represented using a seaborn graph as given in Fig. 4. This can be important in the order that we are able to see the link between a movie's specific rating and therefore how much the movie got. Therefore, we must set a threshold for a minimum number of ratings while constructing a system that recommends. So, to create this new column we use the utility of pandas' groupby. We groupby the title columns, so use the calculation method to calculate the number of ratings each movie received as shown in Fig. 5. The tags for all the movies are combined with the genre to generate a larger metadata for the movies as shown in Fig. 6. This metadata can be used to perform a Content Based approach. The goal is to keep the number of features as low as possible without compromising on the accuracy of the results.

An added benefit of keeping the features lower is that it is less complex when it comes to calculation. A lighter model will help improve the inference time.

We create the value of movie data 'rating' using movie title and calculate rating count in 'title' by applying threshold and get the result.

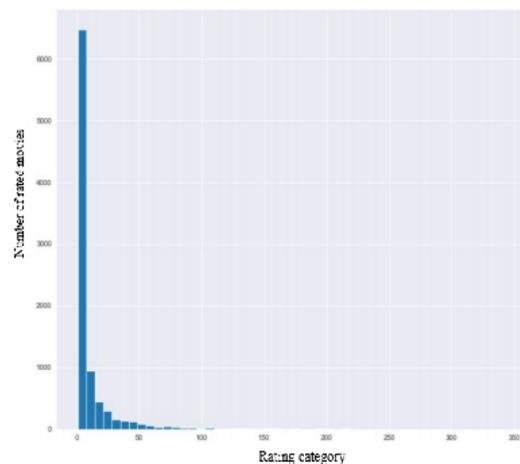


Fig. 3. Number of Rated Movies vs Number of Rated Category.

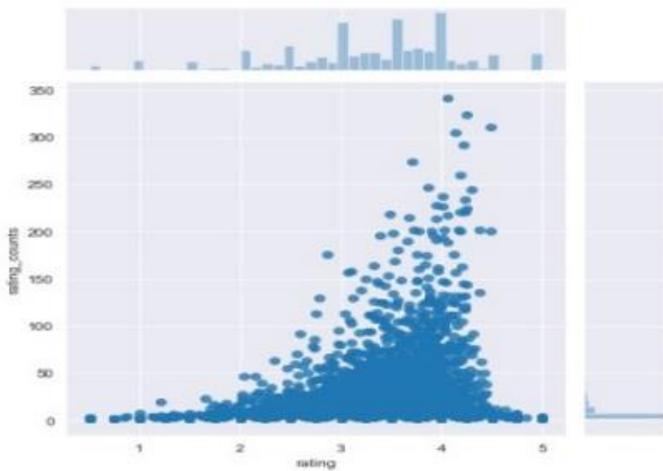


Fig. 4. Graph Representing Retained Movies vs Reduced users.

userid	movieid	rating	timestamp
0	1	1	4.0 964982703
1	1	3	4.0 964981247
2	1	6	4.0 964982224
3	1	47	5.0 964983815
4	1	50	5.0 964982931
...
100831	610	166534	4.0 1493848402
100832	610	168248	5.0 1493850091
100833	610	168250	5.0 1494273047
100834	610	168252	5.0 1493846352
100835	610	170875	3.0 1493846415
...
100836

100836 rows × 4 columns

Fig. 5. Movie Data with MovieId and Ratings.

Index	movieid	Title	Genres
0	1	Toy Story(1995)	Adventure Animation Children Comedy Fantasy
1	2	Jumanji (1995)	Adventure Children Fantasy
2	3	Grumpier Old Men (1995)	Comedy Romance
3	4	Waiting to Exhale (1995)	Comedy Drama Romance
4	5	Father of the Bride Part II (1995)	Comedy
...
9737	193581	Black Butler: Book of the Atlantic (2017)	Action Animation Comedy Fantasy
9738	193583	No Game No Life: Zero (2017)	Animation Comedy Fantasy
9739	193585	Flint (2017)	Drama
9740	193587	Bungo Stray Dogs: Dead Apple (2018)	Action Animation
9741	193609	Andrew Dice Clay: Dice Rules (1991)	Comedy

Fig. 6. Movies with Genres and Tags.

B. Filtering based on Content

The data is sampled to take a large chunk to make a training set on which the SVD loss will be trained. The movie genres are combined with tags to create the metadata of the movies. This metadata will be used to generate a Content Based Recommendation model. A segment of the data is segregated where it contains the user ID, the movie ID and the rating that the user gave to the movie as shown in Fig. 7.

Title	Rating	Rating count
Toy Story(1995)	1.750000	3
Jumanji (1995)	3.84333	2
Grumpier Old Men (1995)	2.00000	2
Waiting to Exhale (1995)	0.60000	1
Father of the Bride Part II (1995)	2.32000	16
Black Butler: Book of the Atlantic (2017)	3.50000	1
No Game No Life: Zero (2017)	5.00000	3
Flint (2017)	3.050632	4
Bungo Stray Dogs: Dead Apple (2018)	2.25000	24
Andrew Dice Clay: Dice Rules (1991)	3.75000	2

Fig. 7. Number of Rating of Movies after Threshold.

This data will be utilized to build the Collaborative model of the hybrid. Additionally, the movie ID and the genres as well as the tags related to the movie are segregated for building the content matrix for the hybrid system. A pass of Singular Value Decomposition is performed in order to flatten the matrix dimensions even further by introducing factorization. This also gives an idea of the variance, which indicates that the first 25 components in the ratings explain the majority of the variance.

This allows us to be even more selective for the data. SVD [22] was chosen as the preferred decomposition method as it gives reliable results and there is some flexibility on how many folds of data we can choose to train on. Furthermore, by our previous review it has been established that it is a good way to ensure high precision.

This adds up in the end when the actual recommendations are generated. Furthermore, TF-IDF is utilized to empower the hybrid recommendation module. This works well with the SVD used earlier.

IV. COMPARISON AND ANALYSIS OF VARIOUS RECOMMENDATION APPROACHES

The Hybrid Recommender System is built with two main components, the Collaborative matrix and the Content matrix. First, the matrix of movies and their ratings are transformed into a feature matrix as given in Fig. 8. This matrix contains the movies against the users and the data contained is the rating given by the user. This featurization is done by utilization of Term Frequency-Inverse Document Frequency. This creates a large number of features, but decomposition will allow these features to be lessened, ultimately bringing down the complexity of the calculations needed. When Singular Value

Decomposition (SVD) is used on this matrix it reduces the features hundredfold. Moreover, analysis after SVD revealed that most of the variance in the data comes from about the first 125 features, which limits the features even further.

userid	1	3	4	5	6	7	8	9	10	11	...	601
movieid												
1	4.0	0.0	0.0	4.0	0.0	4.5	0.0	0.0	0.0	0.0	...	4.0
2	0.0	0.0	0.0	0.0	4.0	0.0	4.0	0.0	0.0	0.0	...	0.0
3	4.0	0.0	0.0	0.0	5.0	0.0	0.0	0.0	0.0	0.0	...	0.0
4	0.0	0.0	0.0	0.0	3.0	0.0	0.0	0.0	0.0	0.0	...	0.0
5	0.0	0.0	0.0	0.0	5.0	0.0	0.0	0.0	0.0	0.0	...	0.0
6	4.0	0.0	0.0	0.0	4.0	0.0	0.0	0.0	0.0	5.0	...	0.0
7	0.0	0.0	0.0	0.0	4.0	0.0	0.0	0.0	0.0	0.0	...	0.0
8	0.0	0.0	0.0	0.0	3.0	0.0	0.0	0.0	0.0	0.0	...	0.0
9	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0
10	0.0	0.0	0.0	0.0	3.0	0.0	2.0	0.0	0.0	3.0	...	0.0

Fig. 8. Feature Matrix with Movies and Ratings.

The algorithm for Hybrid SVD is proposed and it is given with detailed procedure as shown in Fig. 9 for utilizing the concept of standard SVD and enhanced further to acquire the hybrid enhanced method for obtaining low computation time for recommendation procedure.

A. Comparison on Traditional Approach

Root-Mean-Square-Error (RMSE) between real ratings and predictions is a widely used measurement. The lower the RMSE, the more accurately the recommendation algorithm predicts user ratings. To get the initial phase of results on a smaller scale of dataset we decided to use Root Mean Square Error method to find out relevant recommendations as required by the user. Root Mean Square Error [26] method is frequently used method to calculate the difference between the observed measure and the predicted measure. This measurement is usually done using a mathematical formula which is as follows in (1):

where,

n = total number of values present.

p = predicted value.

o = observed value.

i = the value at given position.

So, using this formula and the data from the datasets, recommendations of movies were obtained at initial stage. Here, we used unsupervised learning to classify the data according to our needs from the dataset.

$$RMSE = \sqrt{\frac{\sum_{i=1}^n (P_i - O_i)^2}{n}} \quad (1)$$

We gave some input factors as to get the relevant recommendations.

Algorithm: The HybridSVD-based dimension reduction algorithm

Input: Matrix P

Output: Correlation matrix Q,P,T

1. P is factored into three matrices $Q_1 P_1 T_1$;
2. Reduce the $n \times n$ matrix P_1 to have only k largest diagonal values and obtain a matrix $P_i, i < n$.
3. The matrices Q_1 and T_1 are reduced accordingly, then the reconstructed matrix $S_k = QP_kT$ is the closest rank-n matrix to S, $S_k \approx S$;
4. Calculate the square root of P to be $P^{1/2}$ and calculate the two correlation matrices $QP^{1/2}, P^{1/2}T$.
5. Fill in the unrated data and smooth the matrix. Replace all missing values with the corresponding column average and subtract the corresponding row average from every matrix entry.
6. Calculate the SVD of S according to steps 3 and obtain the matrix Q,P,T.
7. Calculate the user correlation matrix $QP^{1/2}(m,k)$, denoted by U, and the item correlation matrix $P^{1/2}T(k,n)$, denoted by I.
8. For the matrix U(m,k), clustering method is used to obtain the similar rating user-modes and obtain the user-based commendation set P_u .
9. For the matrix I(k, m), a set of top-K similar items towards item i_j will be generated according clustering algorithm. Calculate the prediction for user on page u_a and obtain the item-based commendation set P_i .
10. Sort P_u and P_i by predicted rating and obtain the commendation set P.

Fig. 9. Hybrid SVD Algorithm.

In the experiment, comparing with the traditional UBCF, IBCF algorithm, we can learn that the HybridSVD algorithm can consistently get a lower RMSE and provide better quality of predictions as represented in Fig. 10. The density of a rating matrix can have a significant impact on the performance of collaborative filtering.

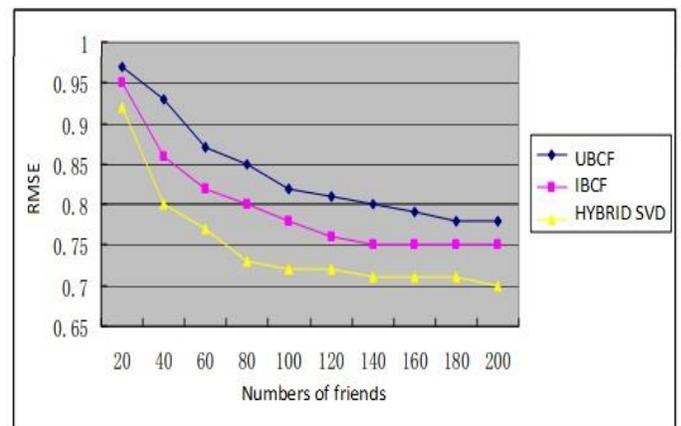


Fig. 10. Comparison of Customary and the HybridSVD Algorithm.

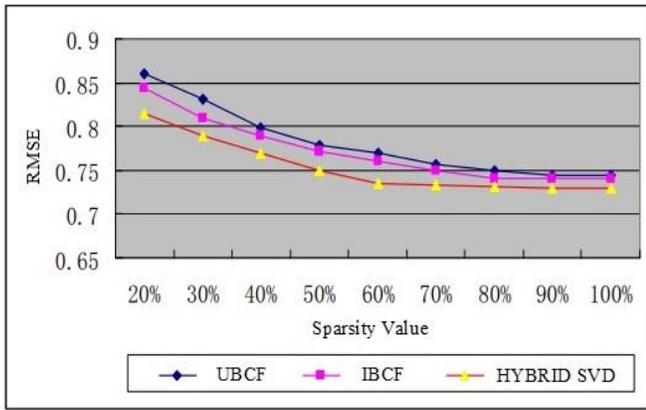


Fig. 11. Traditional vs Hybrid with Sparse Data.

In Fig. 11, we analyze how RMSE evolves with the density of rating matrix. The results indicate that the hybrid approaches consistently improves the recommendation performance regardless of sparsity of test users or items.

B. Sentiment Analysis for Ranking Calculation

The Sentiment Analysis [27] is done over the Large Movie Reviews Dataset. The reviews are categorized in a polar way, so 1 is for a positive review and 0 for a negative. The method used for prediction in the proposed system is a multilayer Perceptron model [14]. Such a model has shown to be effective in predicting sentiment. For each user, the interest movie ratings are used as dimensions to create a vector. The similarity between any two users is determined by the cosine of the angle between the vectors of those two users using the formula as given in (2).

$$\text{Similarity} = \cos \theta = \frac{A \cdot B}{|A| \times |B|} \quad (2)$$

For instance, interest movie ratings of two users are {3.5, 1.0, 4.0} and {2.0, 4.0, 0} respectively. The cosine of the angle between two vectors is calculated as 0.84624085163. This implies that the two users are approximately 84% similar to each other with respect to their interests.

Likewise, the calculation is performed for all users with respect to each other and a similarity matrix is generated. The comparison of different scores for movies based on different filtering approaches is given in Table I.

C. Rating Calculation

The rating calculation for the predicted system with sentiment score is calculated as given by (3). These values are linked with the movie titles and averaged according to the titles as shown in Fig. 12. This allows it to be merged with the data used for the hybrid recommendation. This however also reduces the number of movies drastically as the recommendations available are limited.

$$\text{Rating}_{\text{Sentiment}} = \frac{\sum_1^n s_{m,n}}{n} \quad (3)$$

The predictions generated are averaged to reach a general predicted number for the particular title. Finally, the recommendations from the hybrid system are used to predict the top k movies that would be the most relevant according to

the system. These recommendations also have the predicted ratings attached to the movie titles for each user. These ratings are generated by the similarity matrix between the hybrid recommender.

To reduce the time needed to calculate the final recommendations, the proposed system simply takes these movies and then calculates a final ranking for each movie. The rating from Sentiment is reached by averaging over the number of ratings (n) for the movie across the movie title (m) as given in (4). For instance, a user has selected his interest genre as humour. The similarity points of all opted movies in particular genre, say {10,9,9,8,9}, are listed. The mode is calculated as 9. So, the domain score is 9.

TABLE I. COMPARISON OF SIMILARITY SCORES

Title	CBF	CF	HYBRID SVD
Toy Story (1995)	1.000000	1.000000	1.000000
Jumanji (1995)	0.881076	0.541400	0.722641
Grumpier Old Men (1995)	0.874194	0.448877	0.713540
Waiting to Exhale (1995)	0.827519	0.253080	0.473511
Father of the Bride Part II (1995)	0.254270	0.038402	0.374106
Black Butler Book of the Atlantic (2017)	0.222232	0.437753	0.330236
No Game No Life: Zero (2017)	0.223195	0.431043	0.307581
Flint (2017)	0.224140	0.567751	0.249017
Bungo Stray Dogs: Dead Apple (2018)	0.216449	0.244214	0.378974
Andrew Dice Clay: Dice Rules (1991)	0.215749	0.326542	0.257101

movieId	Title	Sentiment
11	Toy Story(1995)	0.57344
62	Jumanji (1995)	0.376860
33	Grumpier Old Men (1995)	0.532132
44	Waiting to Exhale (1995)	0.985450
25	Father of the Bride Part II (1995)	0.890252
103	Black Butler: Book of the Atlantic (2017)	0.290543
105	No Game No Life: Zero (2017)	0.272899
53	Flint (2017)	0.108863
73	Bungo Stray Dogs: Dead Apple (2018)	0.744440
94	Andrew Dice Clay: Dice Rules (1991)	0.535082

Fig. 12. Predicted Sentiment Score based on user Tags.

The process is repeated for all the preferred interests. Using the scores of the interest domains, rather than the raw input of all users alone, can give us a better similarity and the overall precision shall be increased to a certain extent. This final ranking is reached by adding the averaged sentiment score with the predicted ranking. This allows taking into account the sentiment rating without having to compare it with a huge number of movies. In this way, the impact of the sentiment analysis is still relevant but keeps the sorting of the movies from the larger dataset largely dependent on the output from the hybrid recommendations module. For testing the accuracy of the system, the metrics of precision and recall have been used.

For testing the accuracy of the system, the metrics of precision and recall have been used. These have been used by many other works to indicate how accurate the system is. Precision is the ratio between the True Positives (TP) and the total positives predicted by the system. Recall is the ratio between the True Positives and the total TP with False Negatives (FN). So, Precision gives a measure of how accurate the actual predictions are, while recall gives an idea of how many of the predications are actually being considered. F-Measure gives a great idea of accuracy. For F-Measure to be high, both precision and recall have to be high. Precision, recall and F-Measure all have values between 0 and 1. Equation 1 gives the equation for precision, recall and F-Measure [28]. Fig. 13 shows the performance of the proposed system based on accuracy for validation sets containing 1 million review ratings. Both the Precision and Recall are above 0.7 and this causes the average F-Measure to be 0.93, which is highly competitive with other similar systems as can be seen by the study. Table II shows the performance of the proposed system on different number of ratings.

As the results show the best F-Measure comes from the lower amount of ratings. As the number of ratings increases, precision is seen to increase, while recall gets lower. This causes a lower average F-Measure.

However, the accuracy is still very high. As we can see from the measures, the hybrid system itself takes a lot of time to compute the recommendations. However, the addition of the Sentiment Analysis adds very little time to the overall merged system. So, it is still keeping the time relatively low than if the Sentiment Analysis was used with the total system instead.

We conclude from these experiments that the proposed hybrid algorithm is effective at improving the quality of recommendations and accuracy of the proposed technique improved with sentiment score added.

TABLE II. PERFORMANCE MEASURES FOR DIFFERENT RATING COUNTS

Ratings Count	F-Measure	Recall	Precision
20	0.9304	0.791	0.877
40	0.9334	0.728	0.861
60	0.9147	0.719	0.852
80	0.8012	0.7021	0.811

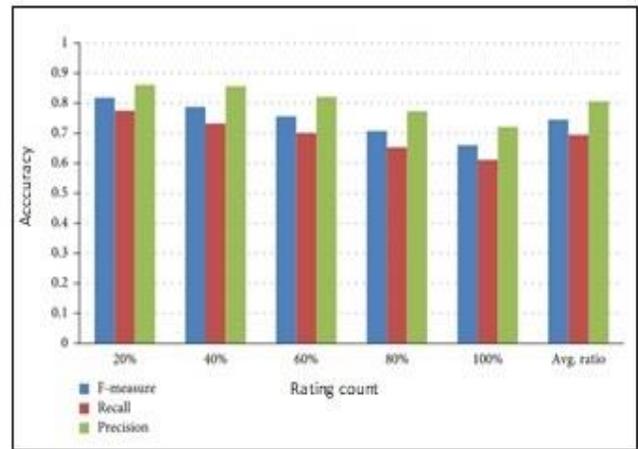


Fig. 13. Predicted Accuracy based on Sentiment Score.

V. CONCLUSION

In this paper a number of studies on recommendation were analyzed and a hybrid recommender system is proposed which works with a Sentiment Analysis model to filter the final results. This system focuses on keeping the computations lesser while still incorporating review data into the recommendations, which contains critical information about the opinions of the viewers. Hybrid SVD is used to generate effective movie recommendations, while a multilayer perceptron is used for Sentiment Analysis to optimize the accuracy level higher. The system performs competitively with other methods, while also incorporating written reviews. For future work, this proposed system might be tested further with a more comprehensive data, for generating recommendations.

VI. FUTURE WORK

The limitation of our work is, we did not have the sentiment score merged with movie dataset. In future research work, we are interested in analyzing the various techniques of sentiment analysis with the respect to the different types of recommendation techniques.

REFERENCES

- [1] Eyjolfsdottir E, Tilak G, Li N (2010) MovieGEN: A Movie Recommendation System. Comput Sci Dep.
- [2] Brusilovsky P, Kobsa A (2007) The Adaptive Web.
- [3] Zhang Z, Zeng DD, Abbasi A, et al (2013) A random walk model for item recommendation in social tagging systems. ACM Trans Manag Inf Syst 4. <https://doi.org/10.1145/2490860>.
- [4] Soni K, Goyal R, Vadera B, More S (2017) A Three Way Hybrid Movie Recommendation Syste. Int J Comput Appl 160:29–32. <https://doi.org/10.5120/ijca2017913026>.
- [5] B.Thorat P, M. Goudar R, Barve S (2015) Survey on Collaborative Filtering, Content-based Filtering and Hybrid Recommendation System. Int J Comput Appl 110:31–36. <https://doi.org/10.5120/19308-0760>.
- [6] Kirmemis O, Birturk A (2008) A content-based user model generation and optimization approach for movie recommendation. AAAI Work - Tech Rep WS-08-06:78–88.
- [7] Zenebe A, Norcio AF (2009) Representation, similarity measures and aggregation methods using fuzzy sets for content-based recommender systems. Fuzzy Sets Syst 160:76–94. <https://doi.org/10.1016/j.fss.2008.03.017>.
- [8] Sarwar B, Karypis G, Konstan J, Riedl J (2001) Item-based collaborative filtering recommendation algorithms. Proc 10th Int Conf

- World Wide Web, WWW 2001 285–295. <https://doi.org/10.1145/371920.372071>.
- [9] A. Patil S, Pagare R (2014) Enhanced Hybrid Recommender System using Social Friend Network. *Int J Manag Inf Technol* 10:2023–2031. <https://doi.org/10.24297/ijmit.v10i4.627>.
- [10] Wang Z, Yu X, Feng N, Wang Z (2014) An improved collaborative movie recommendation system using computational intelligence. *J Vis Lang Comput* 25:667–675. <https://doi.org/10.1016/j.jvlc.2014.09.011>.
- [11] Fang, Z., Zhang, L., & Chen, K. (2016). Hybrid Recommender System Based on Personal Behavior Mining. *ArXiv*, abs/1607.02754.
- [12] Raval T, Patel Y (2017) A Survey: Collaborative Filtering, Content-based Filtering, Hybrid Recommendation Approach. 193–197.
- [13] Ye H (2011) A personalized collaborative filtering recommendation using association rules mining and self-organizing map. *J Softw* 6:732–739. <https://doi.org/10.4304/jsw.6.4.732-739>.
- [14] Jotheeswaran J, Koteeswaran S (2015) Decision tree based feature selection and multilayer perceptron for sentiment analysis. *ARPN J Eng Appl Sci* 10:5883–5894.
- [15] Ashok M, Rajanna S, Joshi PV, Kamath SS (2016) A personalized recommender system using Machine Learning based Sentiment Analysis over social data. 2016 IEEE Students' Conf Electr Electron Comput Sci SCEECs 2016. <https://doi.org/10.1109/SCEECs.2016.7509354>.
- [16] Qian X, Feng H, Zhao G, Mei T (2014) Personalized recommendation combining user interest and social circle. *IEEE Trans Knowl Data Eng* 26:1763–1777. <https://doi.org/10.1109/TKDE.2013.168>.
- [17] Yuan X, Lee JH, Kim SJ, Kim YH (2013) Toward a user-oriented recommendation system for real estate websites. *Inf Syst* 38:231–243. <https://doi.org/10.1016/j.is.2012.08.004>.
- [18] Singh VK, Piryani R, Uddin A, Waila P (2013) Sentiment analysis of movie reviews: A new feature-based heuristic for aspect-level sentiment classification. *Proc - 2013 IEEE Int Multi Conf Autom Comput Control Commun Compress Sensing, iMac4s 2013* 712–717. <https://doi.org/10.1109/iMac4s.2013.6526500>.
- [19] Nazim Uddin M, Shrestha J, Jo GS (2009) Enhanced content-based filtering using diverse collaborative prediction for movie recommendation. *Proc - 2009 1st Asian Conf Intell Inf Database Syst ACIIDS 2009* 132–137. <https://doi.org/10.1109/ACIIDS.2009.77>.
- [20] Bellogín A, Wang J, Castells P (2011) Text retrieval methods for item ranking in collaborative filtering. *Lect Notes Comput Sci (including Subser Lect Notes Artif Intell Lect Notes Bioinformatics)* 6611 LNCS:301–306. https://doi.org/10.1007/978-3-642-20161-5_30.
- [21] Yang D, Zhou Z (2013) Personalized mining of preferred paths based on web log. *Proc 2013 IEEE 11th Int Conf Electron Meas Instruments, ICEMI 2013* 2:993–997. <https://doi.org/10.1109/ICEMI.2013.6743199>.
- [22] Paterek A (2007) Improving regularized singular value decomposition for collaborative filtering. *KDD Cup Work* 2–5. <https://doi.org/10.1145/1557019.1557072>.
- [23] Guia M, Silva RR, Bernardino J (2019) A hybrid ontology-based recommendation system in e-commerce. *Algorithms* 12:1–19. <https://doi.org/10.3390/a12110239>.
- [24] Chen J, Zhao C, Uliji, Chen L (2020) Collaborative filtering recommendation algorithm based on user correlation and evolutionary clustering. *Complex Intell Syst* 6:147–156. <https://doi.org/10.1007/s40747-019-00123-5>.
- [25] Harper FM, Konstan JA (2015) The movielens datasets: History and context. *ACM Trans Interact Intell Syst* 5:. <https://doi.org/10.1145/2827872>.
- [26] Silveira T, Zhang M, Lin X, et al (2019) How good your recommender system is? A survey on evaluations in recommendation. *Int J Mach Learn Cybern* 10:813–831. <https://doi.org/10.1007/s13042-017-0762-9>.
- [27] Soong HC, Jalil NBA, Kumar Ayyasamy R, Akbar R (2019) The essential of sentiment analysis and opinion mining in social media: Introduction and survey of the recent approaches and techniques. *ISCAIE 2019 - 2019 IEEE Symp Comput Appl Ind Electron* 272–277. <https://doi.org/10.1109/ISCAIE.2019.8743799>.
- [28] Martin Ward Powers D (2011) ABC Unconscious Computer Interface View project Autonomous Robotics View Project Evaluation: From Precision, Recall and F-Measure to Roc, Informedness, Markedness & Correlation. 2:37–63. <https://doi.org/10.9735/2229-3981>.

Supervisory Control and Data Acquisition System for Machines Used for Thermal Processing of Materials

Diego Patiño¹, Wilson Tafur Preciado², Albert Miyer Suarez Castrillon³, Sir-Alexci Suarez Castrillon⁴

Electronic Engineering, University Francisco of Paula Santander, Cúcuta, Colombia¹

Faculty of Engineering and Architecture, University of Pamplona, Pamplona, Colombia^{2, 3}

Faculty of Engineering, University Francisco of Paula Santander Ocaña, Ocaña, Colombia⁴

Abstract—A supervisory control and data acquisition (SCADA) system has been developed for three machines used for the thermal processing of materials: a hot wire cutter, an induction heater and a welding test stand. The cutter uses a transformer with adjustable voltage between 20 V and 32 V, and current of 8 A, measuring the temperature of the wire with thermal expansion. The heater uses a 24 V, 15 A sources, and a type K thermocouple embedded in the sample in order to measure temperature. In welding, a temperature control system was implemented for the sample using type K thermocouple and a cooling fan using a 12 V and 20 A sources. The SCADA system consists of a PLC and a PC with a graphical interface which serves to select the process to be worked on as it displays the thermal history of the monitored object. The supervisory system uses a PC with a 32-bit Windows 7 operating system and an OPC software package running on the academic LabVIEW platform. It was designed to use a single human-machine interface for different thermal processes. This paper describes the important components of the system, including its architecture, software development and performance testing.

Keywords—Automatic control of thermal processes; programmable logic controllers; monitoring and supervision in automatic control systems; machine components; Sensors and virtual instruments for control

I. INTRODUCTION

The thermal processing of materials requires keeping track of the measurement of temperatures and their behavior over time, in order to record the thermal history and understand the physical phenomena inherent to each process.

The hot wire cutting machine uses a chrome-nickel alloyed wire between 0.5 and 1 mm in diameter, heated under the Joule effect [1] produced by a source of electrical energy that allows the current to circulate through the wire. When adjusting the operating parameters of the machine, it is always necessary to specify the temperature of the wire and the speed with which the cut must be executed [2]. In addition, the wire must be kept tensed in order to cut certain materials such as foams and Styrofoam [3].

In the case of welding processes, the thermal information of the sample is useful, as it is welded to describe the process phenomena from the point of view of its technology, metallurgy and heat transfer, as required in the development of welding procedures [4]. The measurement of temperatures below the weld bead and its location with respect to the bead is

useful to use a computer program [5], [6], which predicts the temperature distribution of the welded sample. In the particular case of welding materials sensitive to high temperatures, such as manganese-alloyed steels, it is necessary to keep the temperature of the sample below 250°C so as to keep the material from cracking. Fans can be used to cool the sample before continuing to make new welding deposits [7].

The induction heating machine is used for treating ferrous materials with heat [8]. In this case study, small cylindrical samples were heated in order to obtain their heating, maintenance, and cooling cycles. The thermal history experienced by the treated material is of vital importance to learning the behavior and mechanical properties that it will present when used.

In order to be able to interact with the thermal processing machines located in the same laboratory space, a Supervisory Control and Data Acquisition (SCADA) system that supervises and controls the variables present is required. This is achieved with the use of a graphical interface platform such as Labview from National Instruments, and the STEP 7 Micro WIN application for programming the Programmable Logic Controller (PLC) [9]. Other applications have been used for data acquisition in the laboratory and for the control of variables [10]–[13], as is intended with thermal processing machines. No sophisticated temperature control systems were developed [14]–[16], but results were achieved without signal interference.

In this paper, three machines were connected to the SCADA system and, consequently, threads were developed in the PLC to use a single graphical interface from a personal computer (PC). A “Human Machine Interface” (HMI) was designed with control panels for each machine, which can be selected by the user to operate the machines independently.

To describe the SCADA system proposed, the hardware implementations are presented, while the interface developed and the logic applied for the three machines are explained. The three machines used operate independently. These are: the hot wire cutting machine, the induction heating machine and the welding test machine. The rest of the sections in this paper are organized as follows: The next section illustrates the Hardware and Logical interface of the supervisory system. The results are comprehensively presented and discussed in section Results and Discussion.

II. METHODOLOGY

To describe the SCADA system proposed, the hardware implementations are presented, while the interface developed and the logic applied for the three machines are explained. The three machines used operate independently. These are: the hot wire cutting machine, the induction heating machine and the welding test machine.

A. Hardware for the Supervisory System

The system architecture is made up of an HMI interface in charge of supervising the thermal variables of the materials treated and inserting the initial parameters for its execution and a PLC connected to each machine. The PLC is an S7-200 from Amsamotion that has 14 digital inputs and ten digital outputs.

In the cutting machine, PLC controls the movement of the motor and regulates the voltage of the wire, taking the readings of the current and expansion of the wire (Fig. 1). This information is sent to the HMI interface so that the operator can see the thermal behavior of the machine.

The PLC can only apply ON/OFF logic to control the coil in the induction heating machine since it works with a fixed voltage of 24 V (Fig. 2). This machine's interface only shows the temperature of the sample throughout its thermal cycle, that is, the temperature and the time it takes to reach its maximum point and the time it takes to reach ambient temperature.

In the welding machine, the preheating temperature and the temperature at the welding point are monitored (Fig. 3). The functions of the interface for this machine is to set the desired preheat temperature and display the data from the two temperature sensors. This is achieved through two thermocouples connected from the PLC. Additionally, a fan used to cool the sample is directly connected.

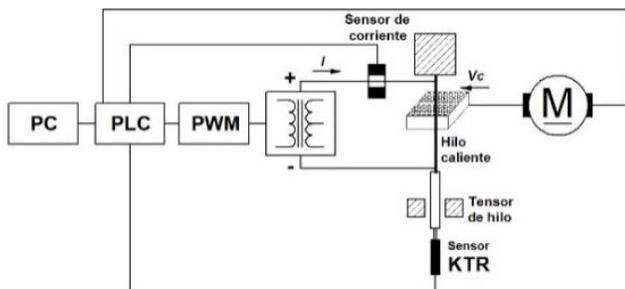


Fig. 1. Structure of the SCADA System on the Hot Wire Cutting Machine.

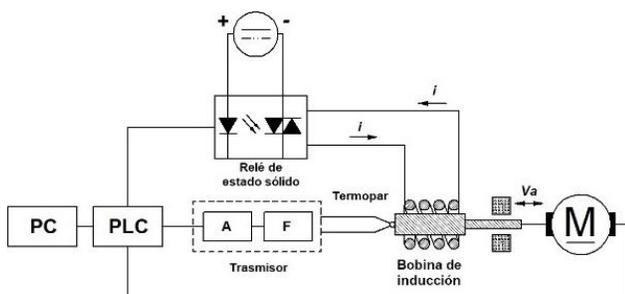


Fig. 2. Structure of the SCADA System in the Induction Heating Machine.

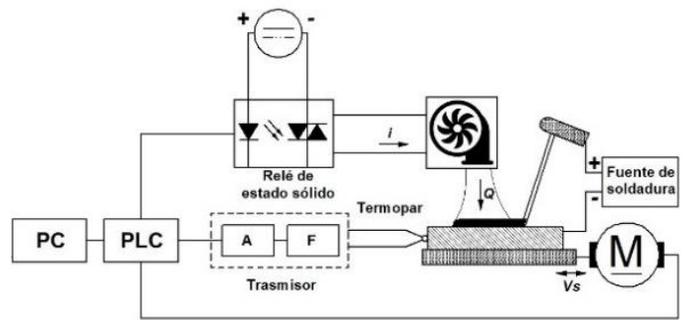


Fig. 3. Structure of the SCADA System in the Welding Machine.

Because the machines are linked to the same PLC, a bar with eight switches was implemented to connect the five sensors (Fig. 4).

It is thus possible to use all five sensors with only two analog inputs. However, only one machine's sensors can be used at a time.

For the operation of the Brushless motor, it was necessary to add a controller that allows manipulating the direction of the motor's rotation and varying the speed. Likewise, to determine the motor's position, Hall Effect sensors were connected to the input pins of the PLC to determine the exact position of the motor's rotor.

For the hot wire cutting machine, a linear expansion sensor and a current sensor are connected to the PLC, manipulated by means of pulse width modulation (PWM). Likewise, the Q0.2 output is connected to an electrical circuit that allows the average voltage flowing through the wire to vary.

Induction heating is commanded by pin Q0.4, which turns the coil on and off. In addition, a type K thermocouple connected to pins 5 and 6 of the switch bar is used, which is in turn connected to the analog input +A of the PLC [17].

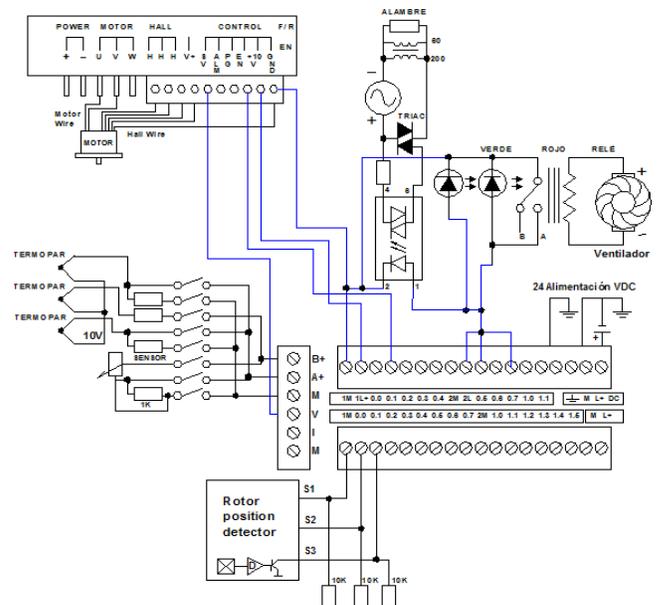


Fig. 4. Structure of the SCADA System in the Induction Heating Machine.

Finally, the pins used for controlling and monitoring the welding machine are Q0.6 and Q0.7, indicating the welding application or the halt of the process. Likewise, pin Q0.8 is responsible for activating or deactivating the fan.

B. Logical Interface of the Supervisory System

For the development of the HMI interface, the LabVIEW computational tool is used. Four virtual instruments (VIs) were implemented: one main instrument (Fig. 5) and three additional ones.

The interface of the hot wire cutting machine (Fig. 6) specifies the configuration of the wire according to its type and dimensions, the maximum current values, temperature, and expansion, in addition to the length of the wire and its electrical resistance. Next, the operating parameters are regulated, namely the voltage of the source and the cutting speed.

A graph is displayed for the current flowing through the wire and another for the temperature measured through the thermal expansion of the wire.

In the interface of the induction heating machine (Fig. 7), the preheat temperature of the workpiece is set. Initially, the sample is heated until it stabilizes at an initial temperature. A ramp then raises the temperature from the initial to the final temperature. In addition, the conditions of the sample are specified, namely initial temperature, length and speed of advance to indicate movement in one direction, in the opposite direction or in alternating direction.



Fig. 5. Main Interface.

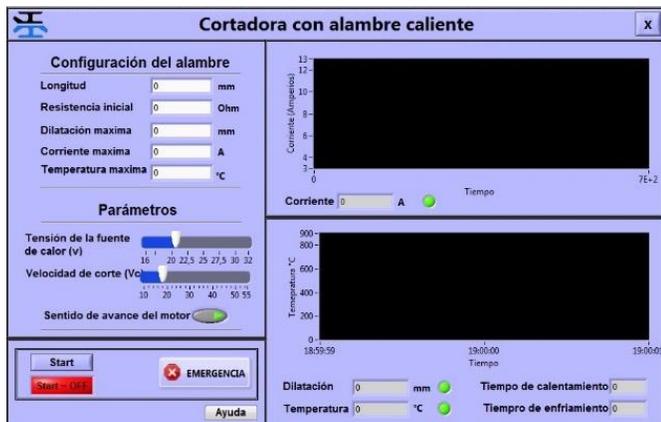


Fig. 6. Hot Wire Cutting Machine Interface.



Fig. 7. Induction Heating Machine Interface.

The temperature that the sample finally reaches is graphically represented to observe the thermal history of the material.

For the interface of the welding testing machine (Fig. 8), the expected preheating and maximum temperatures must be entered, as well as the welding speed for the movement of the sample with respect to the electrode both in forward or reverse directions.

The two thermocouples installed in the sample measure the temperatures below the welding bead and their temperature. These are displayed graphically on the screen and the data is processed for analysis of thermal cycles in welding.

Regarding the software, in addition to the VIs for the interfaces, an automation was designed in the control algorithm for the PLC with the functionalities of the three machines. The PLC programming language used was Ladder, along with the LabVIEW "Object Linking and Embedding for Process Control" (OPC) module for communication between PC, PLC and sensors.

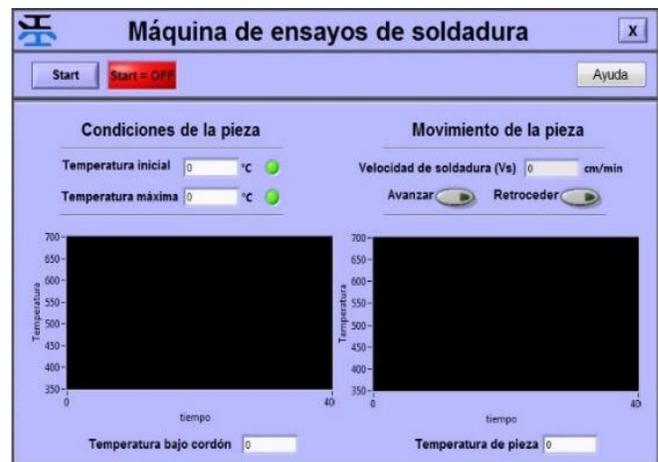


Fig. 8. Weld Testing Machine Interface.

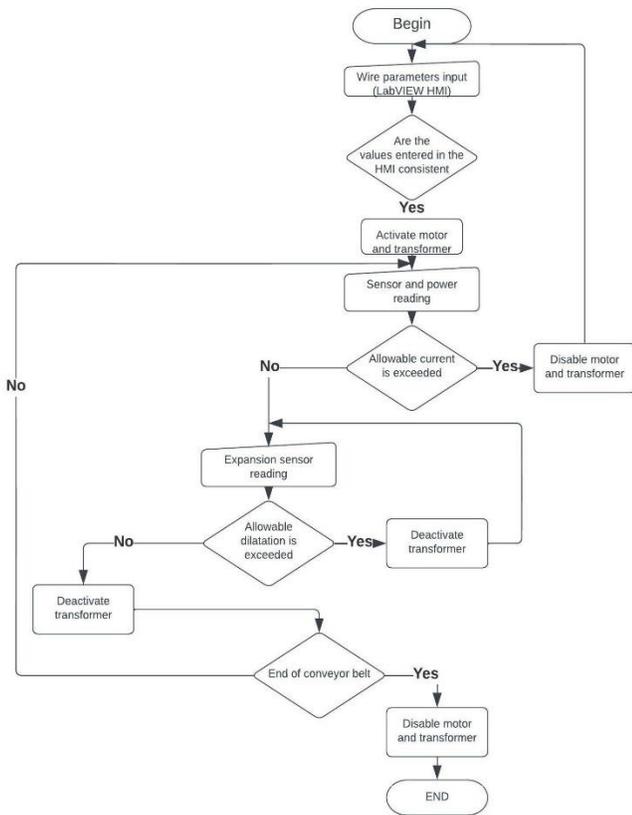


Fig. 9. Flow Chart for Hot Wire Cutting Machine.

The logic of the hot wire cutting machine (Fig. 9) consists of checking the coherence of the entered values, checking the allowed current and the expansion, and finally deactivates the motor and the transformer.

The operating logic for the induction thermal processing machine (Fig. 10) verifies the preheating of the sample and on the other hand the parameters, so as to visualize the temperature ramp with respect to time.

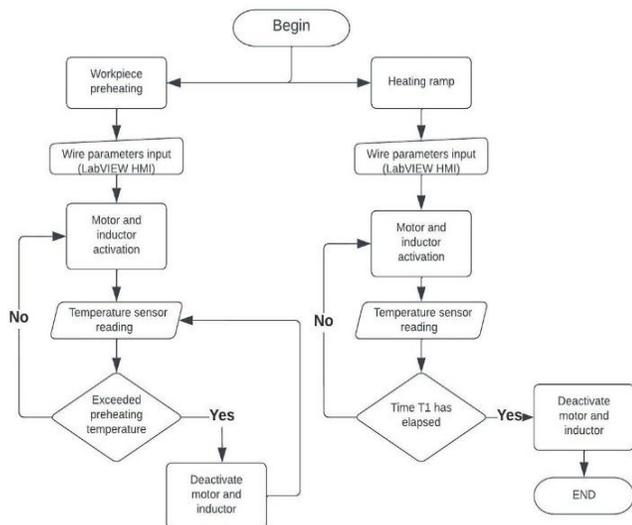


Fig. 10. Flow Chart of Induction Heating Machine.

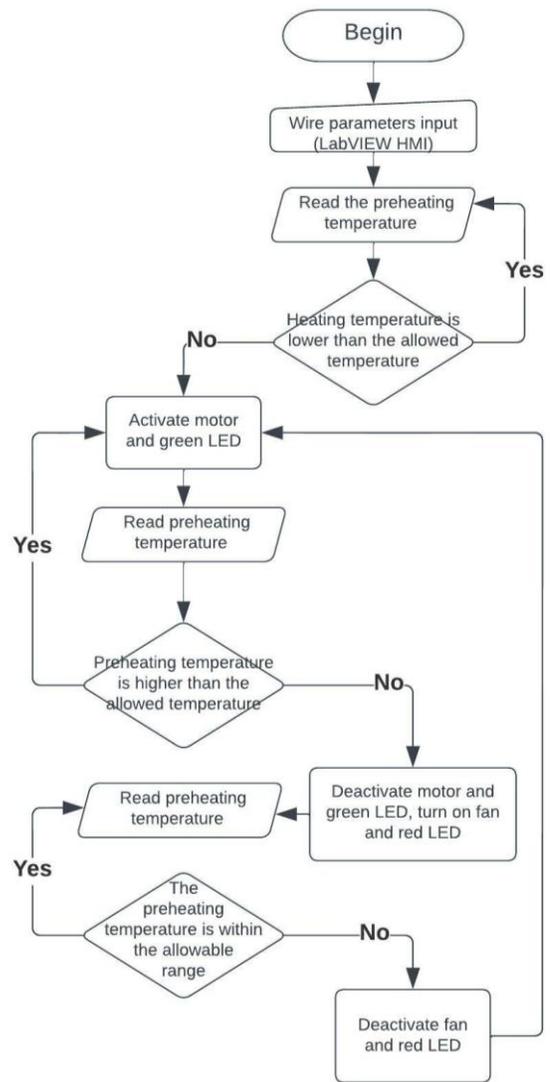


Fig. 11. Flow Chart of the Welding Testing Machine.

The logical process required to work the welding machine (Fig. 11) begins by specifying the preheating temperature of the sample to be welded, delimiting the allowed range. If the temperature is below the range, a red light will go on, indicating that welding cannot start. If the temperature is above the allowed range, it means that the welding must stop and a fan must be activated to cool the welded sample.

III. RESULT AND DISCUSSION

Based on the mechanical design of the machines and the manufacture and installation of components previously carried out by the company Ingeniería Brasilerio Colombiana SAS, the automation of the thermal processes was carried out using the SCADA system. First, a transformer with variable voltage between 18 V to 30 V of alternating current (AC) was designed and installed, controlled from the PLC by means of an Optotriac used as an interface between the controller and the power Triac of the transformer, in order to vary the voltages. The maximum amperage reached was 9A. Fig. 12 shows the implementation of the hot wire cutting machine.



Fig. 12. Implementation of the Hot Wire Cutting Machine.

Then, a KTR-type displacement sensor was implemented. Since it is welded to the counterweight that tenses the wire, it manages to measure the linear expansion caused by its heating. This sensor measures from 0 cm to 2.5 cm, and has a maximum error of 0.02 mm. Measuring the linear thermal expansion of the wire when heated allowed making an indirect measurement of temperature using (1).

$$T_f = T_o + (L_f - L_o) / \alpha L_o \quad (1)$$

Where, L_f and L_o are the final and initial lengths of the wire; T_f and T_o are the final and initial temperatures of the wire; and α is the coefficient of thermal expansion obtained from the wire manufacturer.

In addition to those with variable voltage, tests were carried out using a transformer with a fixed voltage of 20 V. This was done to determine the reliability of the SCADA system function and to compare its accuracy. In this way, it was possible to determine the feasibility of using PWM to manipulate the output and obtain different voltages to heat the wire. Table I shows the comparison between the two voltage sources. A wire with a diameter of 0.7 mm was used for this, varying between three different lengths (800 mm, 900 mm and 1000 mm), as shown. To obtain an acceptable margin of error, three iterations were performed for each length value, with a total of 9 samples for each source.

TABLE I. HOT WIRE TESTS

Length (mm)	Variable source		Fixed source	
	T_{\max} (°C)	I (A)	T_{\max} (°C)	I (A)
800	788	8.7	779	8.6
800	791	8.7	780	8.6
800	790	8.7	784	8.6
900	651	8.2	642	8.15
900	654	8.2	640	8.15
900	649	8.2	640	8.15
1000	583	8	576	8
1000	585	8	578	8
1000	580	8	576	8

When comparing the temperatures measured with each source, a maximum error of 5 °C can be seen. With this error in mind, the use of the sensor can be considered as a good alternative to measure the temperature indirectly in the wire. It is also possible to validate that the results obtained with variable tension are consistent with those obtained with fixed tension.

In the tests with the induction heating machine, an Adeeing Flyback heating plate of 1000 W and 20 A was used for low voltage induction ZVS between 12 V and 48 V. The power supply is carried out with a 24V DC switched source with 15A DC current. The steel sample to be heated must dimensionally comply with a 1/3 ratio between diameter and length [18] for effective induction in thermal processing. Thus, the sample was 15 mm in diameter and 45 mm in length. Simulation tools can also be used to determine the parameters of the induction coil [19]. Fig. 13 shows how the test body reached the bright cherry red required to treat steel with heat. The temperature in these conditions must exceed 700°C up to a maximum of 1100°C. The measurement made with the thermocouple verified that the temperature reached in the process exceeded 700°C. The design of the coil was beneficial for this positive result: it was made of 6 mm diameter copper tubing, obtaining a 7-turn cylindrical shape with an internal diameter of 35 mm and a height of 50 mm. In another phase, the programming of a heating ramp from an initial temperature to a final temperature was tested. This was done by configuring the ignition time of the coil with an ON/OFF control commanded from LabVIEW.

In order to determine the best location of the thermocouple tip with respect to the center of the coil, four tests were performed at distances of 0, 5, 10 and 15 mm.

The thermocouple placed in the middle of the inductor presents a high interference (green line in Fig. 14) due to the magnetic field generated by the coil. When the thermocouple is withdrawn 15 mm, the interference disappears (blue line in Fig. 14). This was confirmed by previous experiments [20].

Finally, the implementation in the welding testing machine was carried out. In it, a car powered by a Brushless motor, previously used in the cutting machine, a press was adapted to hold the sample to be welded. Two K-type thermocouples were inserted into a steel plate to monitor temperatures. For each thermocouple, a transmitter that amplifies and filters the temperature signal to take it to the PLC was used. From the HMI interface, the motor was activated to move along the car with the sample at the welding speed when the electric arc was ignited (Fig. 15).



Fig. 13. Induction Heating Test.

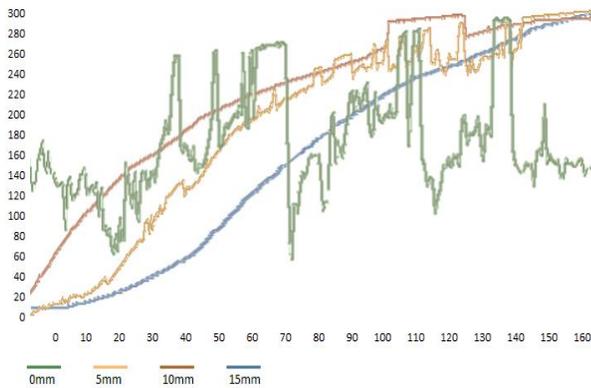


Fig. 14. Preheat Temperature at Different Distances from the Coil Center.



Fig. 15. Implementation in the Welding Testing Machine.

Two lights (one red and one green) are placed on the machine to indicate when the temperature to start welding is reached, which turns on the green light. The red pilot lights up when the sample reaches the maximum temperature previously specified on the HMI. When this happens, a fan that accelerates the cooling of the sample is activated to continue the welding operations.

Additionally, a 12 V and 20 A switching power supply was used to start the fan. The location of the thermocouple tip with respect to the heat source (the electric arc) is also important in order to avoid interference in the signal due to the magnetic field of the welding current, as happens when the thermocouple is moved between 5 and 10 mm away (Fig. 16). A way to measure the temperature in welded plates is thus indicated in comparison with modern techniques [21] and analytical methods [22].

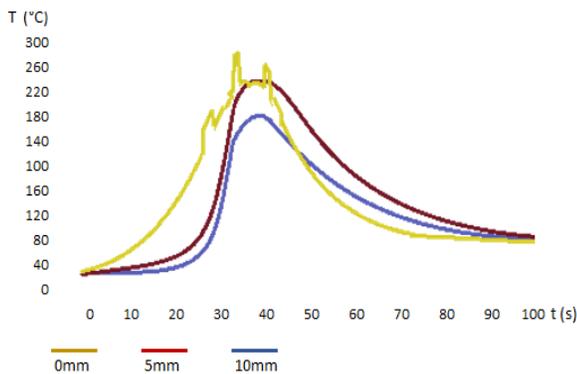


Fig. 16. Temperature at Different Distances from the Weld Bead.

IV. CONCLUSION

In the hot wire cutting machine, the electronic circuit composed of Triac and with the help of the PWM signals generated by the PLC maintains coherence with the linearity of only a maximum error of 1.4%, which allows smooth variations in the voltage, continuously and safely.

In the induction heating machine, it is possible to bring the steel to the temperature necessary for the quenching heat treatment. It is also possible to locate the thermocouple inside the sample, eliminating interference in the signal with a delay of no more than 40 seconds, which could be reduced by increasing the power of the inductor.

In welding, it is important to regulate the advancement speeds of the cord, as well as the measurement of the thermal cycles experienced in the welded area, which was fully accomplished thanks to the implemented automation. To correctly locate a thermocouple on the welded plate, it must be drilled so that when inserting the thermocouple it is spaced between 5 and 10 mm from the center of the weld bead. An online application should be developed to allow remote management outside the laboratory in future work.

ACKNOWLEDGMENT

We thank the company Ingeniería Brasileiro Colombiana SAS for facilitating the laboratory testing infrastructure as well as the thermal processing machines which allowed carrying out this project. We also express our appreciation towards the Francisco de Paula Santander University and the University of Pamplona for their professional and technical support.

REFERENCES

- [1] W. Ma, S. Shi, y X. Zhang, «Three-wire method to characterize the thermoelectric properties of one-dimensional materials», *J. Vac. Sci. Technol. B Nanotechnol. Microelectron. Mater. Process. Meas. Phenom.*, vol. 36, p. 022903, mar. 2018, doi: 10.1116/1.5022118.
- [2] D. G. Ahn, S. H. Lee, y D. Y. Yang, «Investigation into development of progressive-type variable lamination manufacturing using expandable polystyrene foam and its apparatus», *Proc. Inst. Mech. Eng. Part B J. Eng. Manuf.*, vol. 216, n.o 9, pp. 1239-1252, sep. 2002, doi: 10.1243/095440502760291790.
- [3] H. L. Brooks y D. R. Aitchison, «Force feedback temperature control for hot-tool plastic foam cutting», *Proc. Inst. Mech. Eng. Part B J. Eng. Manuf.*, vol. 224, n.o 5, pp. 709-719, may 2010, doi: 10.1243/09544054JEM1717.
- [4] A. L. Voigt, T. V. da Cunha, y C. E. Niño, «Conception, implementation and evaluation of induction wire heating system applied to hot wire GTAW (IHW-GTAW)», *J. Mater. Process. Technol.*, vol. 281, p. 116615, jul. 2020, doi: 10.1016/j.jmatprotec.2020.116615.
- [5] Q. Chen, F. Fei, S. Yu, C. Liu, J. Tang, y X. Yang, «Numerical Simulation of Temperature Field and Residual Stresses in Stainless Steel T-Joint», *Trans. Indian Inst. Met.*, vol. 73, n.o 3, pp. 751-761, mar. 2020, doi: 10.1007/s12666-020-01890-3.
- [6] M. Perić, I. Garašić, Z. Tonković, T. Vuherer, S. Nižetić, y H. Dedić-Jandrek, «Numerical prediction and experimental validation of temperature and residual stress distributions in buried-arc welded thick plates», *Int. J. Energy Res.*, vol. 43, n.o 8, pp. 3590-3600, 2019, doi: 10.1002/er.4506.
- [7] M. F. Ferreira, Â. V. dos Reis, y A. L. T. Machado, «Utilização de revestimento soldado para o aumento da vida útil de sulcadores em semeadoras adubadoras», *Rev. Bras. Eng. E Sustentabilidade*, vol. 2, n.o 2, Art. n.o 2, dic. 2016, doi: 10.15210/rbes.v2i2.8429.
- [8] F. Mühl, J. Jarms, D. Kaiser, S. Dietrich, y V. Schulze, «Tailored bainitic-martensitic microstructures by means of inductive surface

- hardening for AISI4140», *Mater. Des.*, vol. 195, p. 108964, oct. 2020, doi: 10.1016/j.matdes.2020.108964.
- [9] A. Üstündağ y Ç. Gençer, «Designing the Clamp System with the Emergency Braking System in the trains by using PLC and SCADA», en 2021 IEEE 11th Annual Computing and Communication Workshop and Conference (CCWC), ene. 2021, pp. 1436-1441. doi: 10.1109/CCWC51732.2021.9376071.
- [10] M. U. Mahfuz, «Design and Development of a SCADA Course for Engineering Undergraduates», en 2020 IEEE Integrated STEM Education Conference (ISEC), ago. 2020, pp. 1-8. doi: 10.1109/ISEC49744.2020.9280700.
- [11] B. Letowski, C. Lavayssière, B. Larroque, M. Schröder, y F. Luthon, «A Fully Open Source Remote Laboratory for Practical Learning», *Electronics*, vol. 9, n.o 11, Art. n.o 11, nov. 2020, doi: 10.3390/electronics9111832.
- [12] F. J. Maseda, I. López, I. Martija, P. Alkorta, A. J. Garrido, y I. Garrido, «Sensors Data Analysis in Supervisory Control and Data Acquisition (SCADA) Systems to Foresee Failures with an Undetermined Origin», *Sensors*, vol. 21, n.o 8, Art. n.o 8, ene. 2021, doi: 10.3390/s21082762.
- [13] Y. Chen y F. Zhou, «Research and Design of Data Acquisition and Monitoring System for Intelligent Production Line of Prefabricated Building Components», *J. Phys. Conf. Ser.*, vol. 2029, n.o 1, p. 012098, sep. 2021, doi: 10.1088/1742-6596/2029/1/012098.
- [14] D. Muruganandhan, R. Muthunagai, S. Rajkumar, y J. Mohamed Vasif, «Remote Monitoring of Distribution Transformer with Power Theft Detection using PLC amp; SCADA», en 2020 International Conference on System, Computation, Automation and Networking (ICSCAN), jul. 2020, pp. 1-4. doi: 10.1109/ICSCAN49426.2020.9262306.
- [15] K. E. Plewe, A. D. Smith, y M. Liu, «A Supervisory Model Predictive Control Framework for Dual Temperature Setpoint Optimization», en 2020 American Control Conference (ACC), jul. 2020, pp. 1900-1906. doi: 10.23919/ACC45564.2020.9147308.
- [16] E. S. Martynova, V. Y. Bazhin, y V. G. Kharazov, «Increasing the level of control and management of arc steel-smelting furnaces», *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 537, n.o 3, Art. n.o 3, may 2019, doi: 10.1088/1757-899X/537/3/032039.
- [17] D. A. Patiño Epalza, «Diseño de un sistema automatizado para procesos térmicos en la empresa ingeniería brasilero colombiana s.a.s», <http://alejandria.ufps.edu.co/descargas/tesis/1161150.pdf>, 2020, Accedido: 21 de abril de 2022. [En línea]. Disponible en: <http://repositorio.ufps.edu.co/handle/ufps/4506>.
- [18] K. Skalomenos, M. Kurata, H. Shimada, y M. Nishiyama, «Use of induction heating in steel structures: material properties and novel brace design», *J. Constr. Steel Res.*, vol. 148, pp. 112-123, sep. 2018, doi: 10.1016/j.jcsr.2018.05.016.
- [19] H. Sabeeh, I. Abdulbaqi, y S. Mahdi, «Effect of flux concentrator on the surface hardening process of a steel gear», ene. 2018, pp. 80-85. doi: 10.1109/ISCES.2018.8340532.
- [20] H. Kamil, F. Hamdan, F. Abbas, y I. M. Abdulbaqi, «Design and Simulation of a Portable Copper Tubes Induction Brazing Tool for PV System Application», *J. Phys. Conf. Ser.*, vol. 1804, n.o 1, Art. n.o 1, feb. 2021, doi: 10.1088/1742-6596/1804/1/012094.
- [21] N. BarlaDas, P. Ghosh, V. Kumar, N. Paraye, R. Anant, y D. Sourav, «Simulated stress induced sensitization of HAZ in multipass weld of 304LN austenitic stainless steel», 2021, doi: 10.1016/j.jmapro.2020.12.061.
- [22] N. Ma, «Theoretical Prediction of Thermal Cycles and Hardness of HAZ due to Twin Wire Submerged Arc Welding Article in QUARTERLY JOURNAL OF THE JAPAN WELDING SOCIETY • January 2013 DOI: 10.2207/qjws.31.109s READS». 2016.

Modeling Wireless Mesh Networks for Load Management

Soma Pandey¹, Govind R. Kadambi²

Faculty of Engineering and Technology, Ramaiah University of Applied Sciences, Bangalore, India

Abstract—Developing a simulation model for multi-hop multi-gateway wireless mesh networks (WMNs) is a challenging task. In this paper, a multi-hop multi-gateway WMN simulation model is developed in a step-by-step approach. This paper presents a MATLAB Simulink-based simulation model of Wireless Mesh Network (WMN) designed for easy optimization of layer 2. The proposed model is of special utility for the simulation of scheduling of GateWay (GW) and packet within a multi-hop multi gateway wireless network. The simulation model provides the flexibility of controlling the flow of packets through the networks. Load management among the GWs of WMN is performed in a distributed manner wherein the nodes based on their local knowledge of neighborhood beacons optimize their path to a GW. This paper presents a centralized Load Management Scheme (LMS). The LMS is based on the formation of Gateway Service Sets (GSS). The GSS is formed on basis of equal load distribution among the GWs. The proposed LMS is then analyzed for throughput improvement by leveraging the MATLAB Simulink model developed in the paper. A throughput improvement of almost 600% and a 40% reduction in packet loss was observed through simulations thus indicating the efficacy of the proposed LMS. The uniqueness of the simulation model presented in this paper are its scalability and flexibility in terms of network topology parameters.

Keywords—MATLAB; Simulink; multi-hop; wireless mesh network (WMN); gateway; simulation model; load management

I. INTRODUCTION

Nearly always good research methodology is supposed to culminate with performance analysis and simulation study. Research in wireless networks carries no exception. In this paper, a system model is presented to simulate the WMNs. The model is designed in a manner such that it supports the flexibility to increase or decrease the number of GateWays (GWs) within the mesh as well as supports the increase and decrease in the number of hops for a particular router. This results in providing a very simple lightweight model for the optimization of layer 2. This model can be applied easily for GW scheduling in a very quick and efficient manner.

Designing a multi-hop multi gateway WMN model is a tough challenge. In such a mesh network there are multiple parameters to be handled. A very popular example is the IEEE 802.11s mesh architecture [1]. Another example is the Zigbee mesh architecture [2]. In a mesh architecture, the model has to be developed in such a way that the simulations can be performed on a wide variety of scenarios. Specifically, the model should be able to generate scenarios wherein the number of gateways can be varied. At the same time, the model should also have the capability to increase/decrease the number of

Mesh Routers (MRs). The model should also be able to depict the three-level hierarchy of a WMN as presented in Fig. 1 and Fig. 2 of this paper. In this paper, a WMN model is developed which comprises all the aforesaid properties efficiently. Such a model can be used by researchers to apply their solutions easily. The contribution of this paper is to provide a platform and insight which can be used by the scientific community to build their WMN models in a better and more efficient manner using Simulink.

In this paper, a load management technique is applied to reduce congestion in GWs of WMN. In multi-GW WMNs, there is no mechanism available to schedule a fixed set of MRs to GWs. This causes either contention or congestion of GW [3,4]. In this paper, a GW scheduling technique is applied by creating a fixed Gateway Service Set (GSS). GSS is defined as assigning an equal number of MRs to each GW. This will result in avoidance of delay due to repeated computation of nearest GW by MRs. This will also reduce GW congestion by avoiding unfair allocation of MRs to GWs. Such an LMS is then used to analyze the performance of load sharing in various scenarios of WMN. The pertinent point is that the simulations to analyze the proposed LMS could be performed with ease due to the flexibility provided by the proposed simulation model.

II. LITERATURE REVIEW

Most of the research about WMNs generally opts for the ns2 [5] simulator which is an unlicensed open-source simulator. Among the licensed simulators, the top choices are Qualnet and Simulink [9]. Simulations are of two types – Discrete events and continuous ones. Discrete event simulations are suitable for models where parameters do not change until the occurrence of some event. Continuous event simulations are suitable for models where parameters change continuously and do not require the occurrence of any event for the occurrence of change [6]. In [7], the authors have defined and compared different types of simulation models based on event and time. A characteristic feature of a good model is its scalability. The model of WMN should allow flexibility in changing the network topology quickly and easily. An analysis of different types of network simulators is presented in [8].

The model of a WMN comprises three types of nodes – GWs, MRs, and End Nodes/client nodes. GWs provide Internet access to MRs and end nodes. There can be multiple GWs that connect the network to the Internet. The difference between MRs and End nodes is that MRs can redirect traffic to GWs, whereas End nodes just connect to the nearest available MR. Unlike MRs, the End nodes cannot connect to a far-off node

that is out of communication range of a GW. Most of the WMNs follow the generic architecture recommended by the IEEE 802.11s standard [1].

An abundance of literature is available on wireless network simulations using Simulink [9,10,19]. But in most of these papers, the researchers are unable to consider multi-hop multi-gateway WMNs such that the number of GWs is more than 5 and the number of routers more than 20. In [9], the authors have simulated multi-gateway mesh networks. Although the authors of [9] analyze multi-GW association in WMNs, the simulation model comprised only two GWs and 14 MRs. The authors in [10] have simulated WMN with 100 MRs but have not considered more than one GW. This limits their analysis to WMNs with only a single GW. In [11], the authors have used the ns2 simulator to simulate a network with 50 nodes but with only a single GW. The authors in [12] have considered WMN with up to 5 GWs and 50 MRs. But the topology of WMN has not been modified throughout the simulation process.

In this paper, a step-by-step process is presented to develop a WMN on Simulink. Thereafter, the model is used to analyze the performance of the WMN when there is an increase in the number of GWs and MRs systematically. In the process, there is an attempt to achieve a balanced approach for ensuring a fair allocation of MRs among multiple GWs. Therefore, to each GW an equal number of MRs is assigned. The uniqueness of the model presented in this paper is that it is scalable and flexible in terms of network topology parameters. In this model, very easily the number of MRs can be increased and decreased. The same applies to GWs wherein the number of GWs can be increased and decreased with ease.

Finally, this model is applied to analyze a fair GW scheduling technique. This paper proposes to allocate a fixed set of MRs to a GW thus forming a GSS. The GWs are deployed such that each GW receives an equal number of MRs which is in its wireless coverage area. These MRs assigned to a GW might be directly within its communication range or through another MR as mesh topology [1]. This will result in the mitigation of delay in the computation of the path to a GW. Since most of the traffic is to or from the Internet, GWs become a major source of traffic within WMN. This results in congestion of the wireless links to GWs and congestion due to queuing delays in GWs. This occurs since some GWs might have a larger number of MRs associated with them compared to some other GWs in WMN, which might be idle due to no traffic through them. In [13], the authors have identified that the available capacity of a GW reduces by $O(1/n)$ where n is the number of users. The presence of many users results in GW bottleneck and was termed as a 'bottleneck collision domain' in [13]. In [14,16] the authors have attempted mitigation of GW bottleneck with a proposal of performing cooperative caching among the MRs to avoid packet loss. Most of the load balancing techniques focuses on route optimizations. Since most of the time routes are formed on basis of the information about the neighborhood of a node, these are not very optimal techniques. It is also observed by [15] that such techniques consume network resources and bandwidth due to repeated attempts for route requests and route response packets. A very recent paper in this direction is by [17]. In this paper, the authors have attempted to balance load by optimizing the

Adhoc On-Demand Distance Vector (AODV) Routing Protocol. AODV is a very popular routing protocol in WMNs. This is because most of the optimization techniques in WMN are derived from the adaptations of MANET technologies. In the case of MANETS, assigning a fixed service set to GW is not possible due to peer-to-peer routing. But in the case of WMNs, especially IEEE802.11s WMNs, such an assignment is possible because WMNs are a hybrid of fixed and mobile infrastructure. Therefore, this paper proposes a mechanism in which each GW can be assigned its own GSS thereby reducing the network traffic for route determination. This technique also results in a fast and efficient delivery mechanism because the MRs do not have to keep computing the shortest path to a GW when there are multiple GWs in the vicinity. In [18] the authors have performed a detailed survey and have concluded that the WMNs are here to stay for a long time. In [18] the authors also note that the WMN model should become more and more flexible in a manner that allows for the nodes to move from one gateway to another seamlessly. This paper attempts to provide such flexibility while performing the load sharing among gateways through a fair assignment of MRs to GWs.

III. NETWORK ARCHITECTURE OF WMN

Fig. 1 presents a conventional WMN with three levels of nodes. For more details on WMN and its architecture one may refer to [19] in which authors have explained in detail different types of WMNs based on various IEEE standards. At the first level is the GW node which connects to the Internet. In IEEE802.11s standard [1], it is called the Mesh Portal Point (MPP). But in this paper, a more generic term called the GW node is used. The GW node is connected to many routers which are called the Mesh Points (MPs) in IEEE 802.11s. But to keep the term generic, it is called the Mesh Router (MR). Finally, the end nodes are the client devices, for example, a laptop, smartphone, or a sensor. These are called the Mesh Clients in IEEE 802.11s. In this paper, they are referred to as End nodes.

The WMN architecture in Fig. 1 is extended further to a real-world scenario in Fig. 2. The difference between Fig. 1 and Fig. 2 is that in Fig. 1 none of the GWs have a fixed Gateway Service Set (GSS) whereas in Fig. 2 each GW has been assigned a fixed number of MRs.

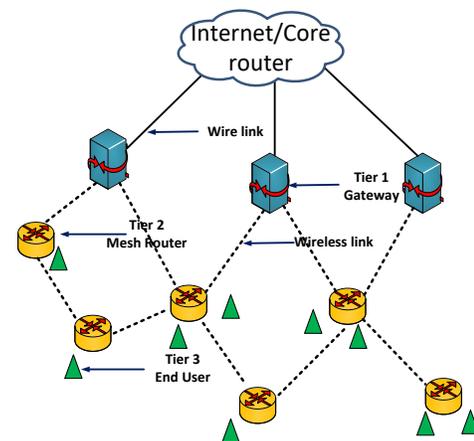


Fig. 1. Three Tier WMN Architecture.

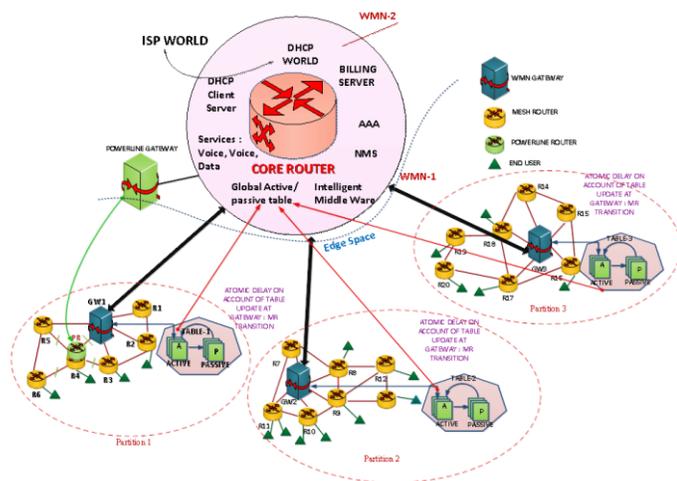


Fig. 2. Schematic of a Real-World WMN Scenario.

In Fig. 2, the MRs and the GWs are connected through wireless links and there are wired connections between the GWs and the core router. Each GW locally maintains an Active Passive Routing (APR) table which is explained later in this section. It is important to note that there might be many WMNs connected and served by the same Internet Service Provider (ISP). To illustrate this, there are two labels indicating WMN-1 and WMN-2 in Fig. 2. Both the WMNs are shown connected to the same ISP and core router. Fig. 2 depicts a dashed line that separates the ISP space and the edge space. The edge space marks the beginning of the WMN space or the Local Area Network (LAN) space.

The next section presents how this model can be used to implement fair scheduling of GWs within a WMN. The proposed scheduling of GWs through the judicious management of the routing table is called Load Management Scheme (LMS). The usage of the term LMS is justified because the scheduling of GWs aims at the reduction of processing load on GWs. A particular GW can hand off its extra load to a neighboring GW by simply changing the MR entry from active to passive, in its routing table. At the same time, the GW receiving the load will change the entry of the MR from passive to active. Therefore, by managing and changing the routing table, a fair allocation of GW to MRs can be maintained.

IV. COMPONENTS OF WMN

The core router is a router that connects the edge network to the Internet. The major functionality of the core router is to streamline the Internet traffic as per the bandwidth demand without a loss in performance [16]. In this paper, the core router is entrusted with the main functionality of meeting the bandwidth demand at the edge routers (GWs in this case). Fig. 2 depicts the Core Router sub-block. The major components of a core router are Internet Service Provider (ISP), Core Router, Authentication, Authorization, and Accounting (AAA) server, Billing server, Network Management System (NMS) Server, Voice, and Video data services, DHCP Server, Global Active/Passive Routing (APR) Table and Intelligent Middle Ware (IMW).

Although the APR Table and IMW may not be a part of the core router in conventional WMN, the proposed LMS requires this additional software to be installed at the core router. The IMW and its components are discussed in more detail in a later section. The components of Edge Space define the components of WMN and are described next.

A. GW

This is the connection point to provide Internet connectivity through the Core router. The GW is connected to the core router through a wired medium or high bandwidth.

B. MRs

The GW is associated with a set of MRs labeled from R_1 to R_6 . Each MR has a wireless connection to either another MR or the GW. An MR is connected to another MR or GW if they are within transmission range of each other. If an MR is not in the transmission range of the GW, it can still reach the GW through multiple hops by using another MR which is within the transmission range of GW. An MR is not allowed to send a packet to GWs other than its associated GW even if there are other GWs that are in its transmission range. The MR-GW association is decided through the APR table depicted alongside each GW in Fig. 2.

C. Local APR Table

This is the most important data structure of the proposed LMS residing at the GW. This table facilitates mapping the GSSs of WMN. Each GW has two main columns in its APR table namely, the active column and the passive column. The active column contains the list of all those MRs that are associated with the GW whereas the passive column keeps the list of all those MRs that are in the transmission range of the GW but are not associated with it. While configuring the WMN, the GW routing table is configured such that the GSSs obtained from the Greedy Graph Partitioning Algorithm (GGPA) are mapped onto the routing table of each GW. This means all those MRs that belong to a GSS are made active in the routing table of GW.

Table I is the routing table of the GW belonging to GSS 1 depicted as partition 1 in Fig. 2. In Table I, the active column contains MRs labeled R_1 , R_2 , R_3 , R_4 , R_5 , and R_6 . The MRs labeled R_7 and R_{11} are associated with GW_2 as shown in Fig. 2. But because they are within the transmission range of GW_1 , R_7 and R_{11} are listed under the passive column of Table I.

TABLE I. APR TABLE FOR GSS 1 (PARTITION 1 IN FIG. 2)

Active	Passive
R_1	R_7
R_2	R_{11}
R_3	
R_4	
R_5	
R_6	

When a GW gets overloaded, some MRs must be offloaded to a neighbor GW. This process becomes remarkably simple with the proposed APR table of the GW. If GW2 becomes overloaded, then it can offload MRs labeled R7 and R11 to GW1 by changing the APR table entry of GW1 and GW2.

TABLE II. APR TABLE OF GW OF GSS1 AFTER RECEIVING MRs FROM GSS 2 FOR LOAD SHARING

Active	Passive
R ₁	
R ₂	
R ₃	
R ₄	
R ₅	
R ₆	
R ₇	
R ₁₁	

Initially, the APR table of GW1 appears like Table I and after offloading of MRs labeled R7 and R11 to GW1, the modified APR of GW2 is depicted in Table II. Interestingly, the transition process at the core router involves only updating the entry of the APR table of receiving and sending GWs.

D. The Intelligent Middle Ware

The core router shown in Fig. 2 executes an IMW explained in this section. The IMW residing on the Core router of Fig. 2 has two modules namely load monitoring and load sharing. The load monitoring module periodically estimates the load demand of each GW. Based on this estimation, the module decides whether the load on GW is excess or not. The basis of this calculation is based on comparing load demand to the capacity of GW. A discussion on computing the capacity of GW is presented in [16,20]. If the load demand exceeds the capacity of GW, then a particular GW is overloaded. Demand at each GW is computed by the load monitoring module by recording the number of MRs connected through the GW and the applications that they are executing.

The load sharing module has two components to support load sharing with neighboring GWs. Load sharing with wireless GWs is invoked when the load monitoring module raises an overload alert. The load sharing module checks the stability condition (whether any neighboring GW is having less load and is willing to receive MRs from the overloaded GW). If the stability condition is satisfied, then from an overloaded GW, an MR with a high bandwidth demand is shifted to a neighboring GW with a nominal load. Accordingly, the APR tables are updated.

V. SIMULATION MODEL OF THE PROPOSED LMS OF WMN

Since this research is focused on routing optimization for load balancing, this study requires simulation to be built around the multi-hop routing mechanism of mesh. In this paper, the Simulink blocks are developed to suit the requirement of IEEE 802.11s MAC which is the main module of interest to this paper. To develop this model, the network architecture evolved in the previous section is used. The

coming sections explore this process and after the explanation of the individual blocks, the complete integrated model is presented. Fig. 3 represents the sub-blocks required for the implementation of the proposed LMS. The simulation model has the following four major modules namely Packet Generator, Core router, MR Mobility, and MPP GW.



Fig. 3. Simulation Models of Sub Blocks for LMS.

These blocks follow the network architecture depicted in Fig. 2. It can be recalled that in the network architecture of Fig. 2, the core router routes packets to the respective GWs. Thereafter the GWs forward the packets to the destined MR through multiple hops. Finally, the destined MR forwards the packet to the client/customer device. The simulation model uses a packet generator to generate traffic and this traffic constitutes the input to the core router. Each packet has a destined MR identified by an IP address. The core router module maintains a global APR table for each of the GWs of WMN.

The MR-mobility module simulates the transition of MRs from one GW to another. Since the MR transition effect percolates to the core router, this block is kept located between the GW and the core router block in the hierarchy. For this, the MR mobility block receives signals from the GW about the completion of the transition. It must then send a signal to the core router to resume packet flow from the packet generators, in case the MR transition is complete. The core router in turn issues a signal to the packet generator to resume packet generation.

The MPP GW module defines the operations of GW. The GW model also has a sub-model that simulates the end-user/customer device which operates on the normal IEEE 802.11x [21] standard. This simulation also comprises the client hand-off/handover process between two MRs. A detailed discussion of the layout of each block is presented in the coming sub-sections. The block layout is followed by a screenshot of the actual Simulink model used in the simulation. More details on the process of designing the Simulink block are presented in [22].

A. Packet Generator Module

A packet generator is modeled as a random packet generator that generates packets with labels for specific GWs. The packet generator has five major blocks as depicted in Fig. 4, namely, Set Attribute, Free running counter, Time based entity generator, Repeating Sequence Stair, and FIFO Queue.

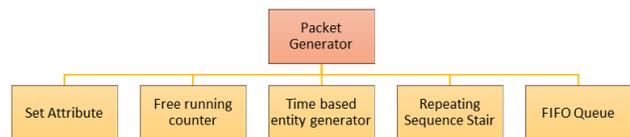


Fig. 4. Block Hierarchy of the Packet Generator Model.

Set attribute block gets packet generated. The free-running counter block is for data generation. Repeating sequence block is to define the length of the data block. Time-based entity generator defines the rate at which packets are generated. Finally, packets are delivered to the core router through the FIFO queue. A screenshot of the final packet generator Simulink block comprising these sub-blocks is shown in Fig. 5. The output of the packet generator is fed to the core router. This module is explained in the next subsection.

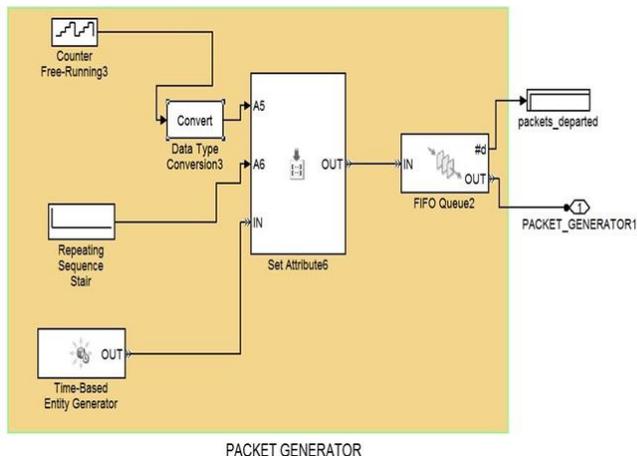


Fig. 5. Simulink Model of Packet Generator.

B. Core Router Module

The function of a core router is to route the packet coming from the packet generator to respective GWs. From its global APR table, the IMW checks for the destination MR and the GW associated with the destination MR. Then the packet is routed to the associated GW of destination MR. The core router has six major blocks as depicted in Fig. 6.

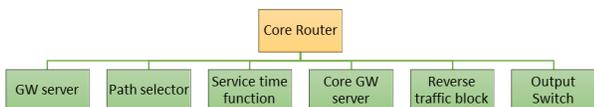


Fig. 6. Hierarchy of the Core Router Block.

1) *GW server* receives packets from respective ‘packet generators’ which are to be fed to ‘path selector’.

2) A *reverse traffic block* is used to model traffic from GWs and MRs to the core router.

3) *Path selector* sorts all the GW packets along with power line packets and reverses traffic packets.

4) *The service time function* computes the time taken for path selection (routing) to respective GWs. The referred service time is computed by dividing the length of the packet by the capacity of the core router. For example, if the length of the packet is 512 bits and the capacity of the core router is assumed to be 100 Mbps, then the service time for each packet is 512 ns.

5) *Core GW server* serves each packet in-service time defined by the service time function. The health of the core router is monitored in the “core GW server” by analyzing

Channel Utilization, Average delay, Number of packets departed, Number of packets dropped, Average wait time, and packet delay.

6) *Output switch* sends packets to their destined GWs: The functionality of the Core router block along with these components is depicted through a Simulink model shown in Fig. 7.

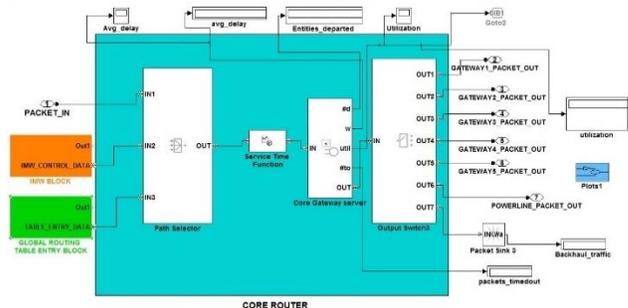


Fig. 7. Simulink Model of Core Router.

C. MR-Mobility Module

MR Mobility module simulates the MR hand-off and hand-over from one MR to another MR just like client mobility in an IEEE 802.11x network. When an MC moves out of an MR access area, the MR hands over the MC traffic to the MR which is accessible to the MC. This delay is added to the total packet response time. A Simulink model of MR mobility is shown in Fig. 8. It can be observed that the MR mobility module has two major blocks.

- MR delay model.
- GW Channel model.

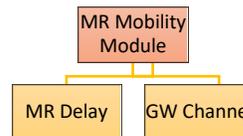


Fig. 8. Functional Blocks of MR Mobility Model.

1) *MR delay model*: The MR delay block simulates delay due to the transfer of MR from one GW table to another. This is an atomic process involving a delay in updating the global APR table at the core router and the local APR tables at the sending and receiving GW involved in MR transition. This block consists of a packet generator sub-block as shown in Fig. 9.

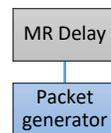


Fig. 9. Sub-blocks of the MR Delay Block.

The packet generator block induces reverse traffic from the GW towards the core router. In this paper, the word ‘traffic’ indicates the traffic from the Internet (packet generator) to the MRs whereas ‘reverse traffic’ is defined as packets from the

MR towards the Internet. This reverse packet, when generated by the packet generator module of the MR delay block, indicates invoking of load sharing. When the reverse traffic packet reaches the main packet generator module through the core router, it blocks the packet generator till the time the MR transition takes place (delay in updating the local and global APR table). Once the MR transition is over, another reverse packet is generated from the MR delay block towards the core router and the packet generator resumes the forward traffic. Fig. 10 depicts the MR delay block comprising the packet generator sub-block.

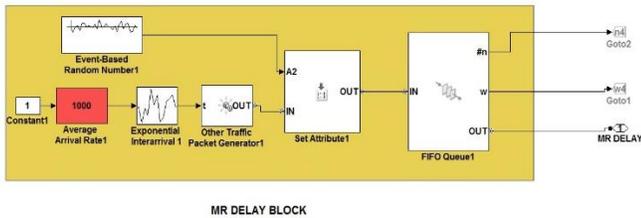


Fig. 10. Simulink Model of MR Delay Block of MR Mobility Module.

The packets generated by the MR delay block are forwarded to the ‘GW Channel Model’, explained in the next section, which incurs channel processing delay.

2) *GW channel model of the MR mobility block:* The GW Channel block of the MR mobility module models the channel between the GW and Core router. The sub-blocks of the channel model are depicted in Fig. 11.

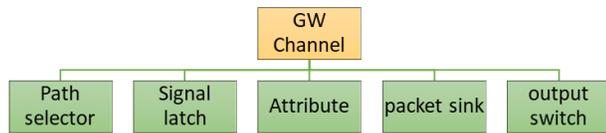


Fig. 11. Sub Blocks Comprising the Channel Model Block.

a) *Path selector:* This module is used to receive packets from the core router (in case of forwarding traffic) and MR delay block (in case of reverse traffic).

b) *Signal latch:* Path selector uses ‘Probabilistic signal latch model’ to select packets.

c) *Attribute:* The service time in processing packets in a single server is calculated by the ‘Attribute Function’ block.

d) *Packet sink:* The data packets are routed to the respective GW or core router in case of backward traffic using the ‘Packet sink’ block.

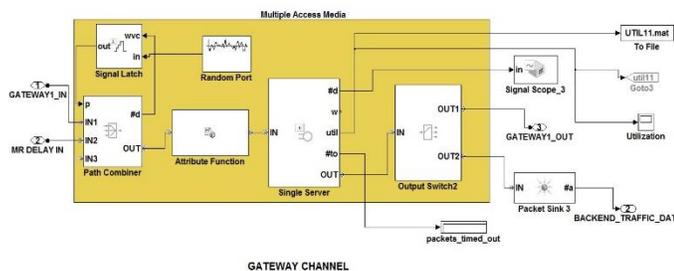


Fig. 12. Simulink Model of GW Channel Block of the MR Mobility Module.

e) *Output switch:* The packets are sorted by ‘Output switch’ which checks the packet header and routes the packets to the respective destination GW or routes the packet back to the core router if it is a backward packet requesting a traffic block. Fig. 12 depicts the final Simulink block of the channel model representing all the above-listed sub-blocks.

3) *The final MR mobility Simulink block:* The final MR mobility Simulink block comprising its major sub-blocks the MR delay block of Fig. 10 and channel model of Fig. 12 is shown in Fig. 13.

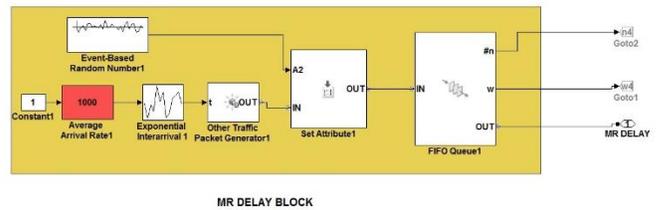


Fig. 13. Simulink Model of MR Mobility Module.

D. GW Module

The GW module segregates and routes the packet destined to a particular MR. The structure of the GW module is depicted in Fig. 14.

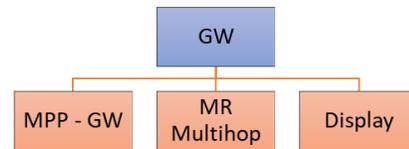


Fig. 14. Hierarchy of the GW Model.

As depicted in Fig. 14, the GW Model has two major blocks:

- MPP GW
- MR Multi-Hop (Multi-Hop of MR)

1) *MPP-GW sub block of the GW:* MPP-GW model further comprises the sub-blocks shown in Fig. 15.

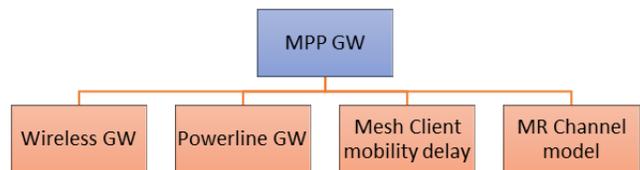


Fig. 15. Sub Blocks of the MPP-GW Block.

This block has four major sub-blocks:

a) *Wireless GW:* This model has a ‘set attribute’ block that receives packets from the core router and forwards them at a pre-defined data rate (bandwidth) to the channel model through the ‘FIFO queue’ block.

b) *Power Line GW:* The power line GW gets the packets directly from the core router. It uses the same ‘set attribute’ block to define the data rate. The ‘FIFO queue’

block gets input about the MRs chosen to transit to the power line. This block is attached to the core router and it routes the traffic from the core to the MRs specifically attached to the power line. The MRs attached to the power line GW can be identified using the global APR table at the core router.

c) *Mesh Client Mobility delay*: This block has two parts. One part simulates the end-user device, and the other part simulates the mobility of the client. The end-user device acts as either a source or a sink for the traffic flow. Therefore, the end-user device is modeled as having two parts- the packet generator part which is the source, and the sink which consumes the packets which are received through its connected MR. The mobility of the client is modeled by introducing a handoff-handover delay whenever the mobile client moves from one MR transmission range to another.

d) *MR Channel model*: This model simulates the channel between the GW and the MRs. The channel model is responsible to induce the client's mobility delay. This is the delay involved in hand-off and hand-over when an end-user moves from one MR to another MR. This is different from the MR mobility delay which involves delay incurred during the transition of MRs for load sharing.

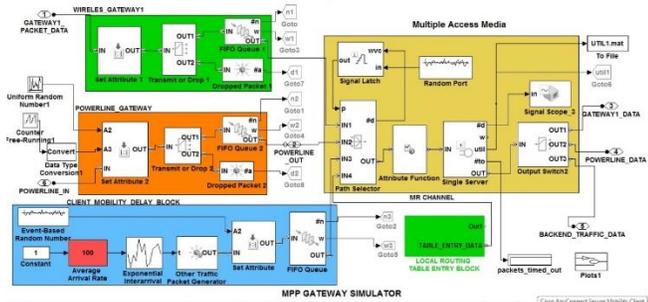


Fig. 16. MPP-GW Simulation Sub Block of the GW Simulator Block.

The sub-blocks listed in Fig. 15 are shown in the final Simulink block of the MPP-GW model block in Fig. 16.

2) *Multi-Hop MR of GW*: The ‘multi-hop MR’ sub-block receives packets from the MPP-GW block and records variation in-service time due to the increasing number of hops. This block consists of two sub-blocks as depicted in Fig. 17.

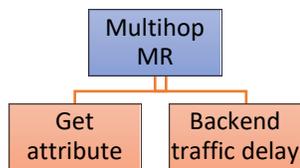


Fig. 17. Sub Blocks of the Multi-hop MR Block.

These blocks are:

a) *Get attribute*: It refers to MR where it receives service time from the respective GW. The service time of the packet will be almost doubled for every hop.

b) *Backend traffic delay*: This block simulates the processing delay incurred by the mesh management traffic. Although this traffic does not contain the actual data packets, mesh management packets are also important. The Simulink block of multi-hop MR simulation is shown in Fig. 18 which depicts the ‘get attribute’ and ‘delay’ sub-blocks listed in Fig. 17.

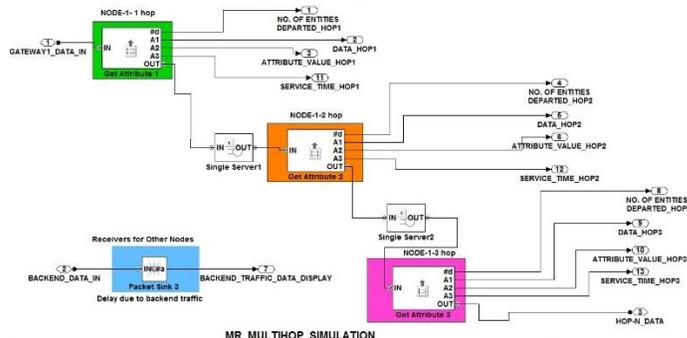


Fig. 18. Multi-hop Simulation Sub Block of the GW Simulation Block.

3) *Display block*: Since this is the final and the last block of the simulation model, it displays parameters such as ‘number of packets departed’, ‘Average waiting time’, and ‘service time’.

E. Final GW Simulation Block

Fig. 19 depicts the final GW simulation block comprising the MPP-GW and the MR multi-hop simulation blocks.

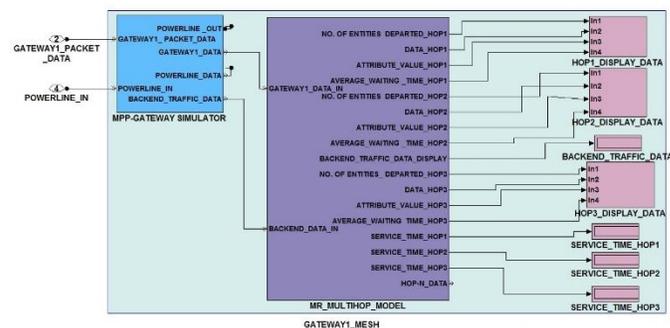


Fig. 19. GW Simulation Block.

F. System-Level Simulation Block

The system-level simulation model is shown in Fig. 20. This model depicts the packets generator module, the core router module, the MR mobility module, and finally the GW module. The process flow can be mapped onto the final simulation model in Fig. 20. This model will be used in the next section to derive various performance results and to investigate the various system parameters. The framework of this model follows the schematic architecture proposed in the network architecture diagram in Fig. 2. The next section presents the results obtained through the system-level simulation block of LMS shown in Fig. 20.

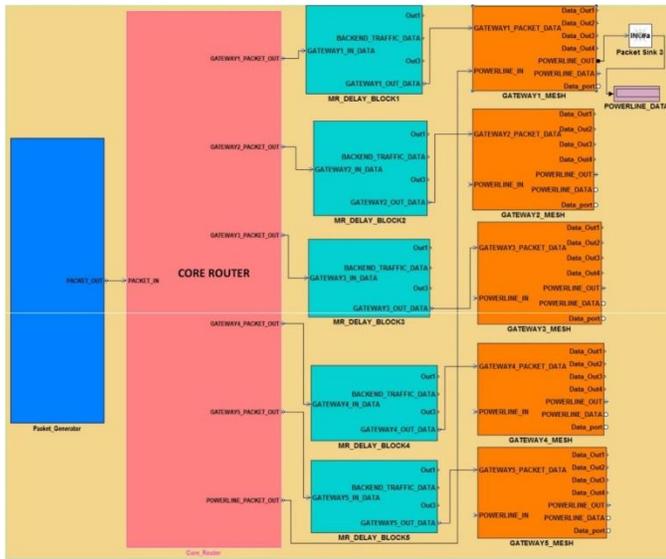


Fig. 20. System-Level Simulation Block of LMS.

VI. ASSUMPTIONS AND PARAMETERS FOR MATLAB MODEL

In the previous sections, a simulation model of the proposed LMS was developed using MATLAB, and Simulink blocks. This model is used to analyze the performance of the proposed LMS. Various test cases for performance analysis are created to compare the performance of the WMN with the proposed LMS and the conventional WMNs with no load management feature. Throughout the simulation process, the simulation parameters are chosen as per Table III.

The traffic flow is assumed to be Markov distributed. The core router is assumed to be connected through a high-speed wire link. Therefore, its capacity is 100 Mbps which is the usual capacity of a high-speed fiber backhaul. The communication range of 250 m and carrier sensing range of 550 m is the most used value in the simulation of the IEEE 802.11 standard. This assumption is based on the ns2[5] and Qualnet’s physical layer adaptation of the IEEE 802.11 standard and 914 MHz Lucent WaveLAN DSSS wireless card [23].

TABLE III. SIMULATION PARAMETERS

Parameter	Value
Number of GWs	1 - 15
Number of MR	Varying (100 – 300)
Maximum number of Mesh Clients	250
Mean Packet arrival rate	0.01s (100 packet/s)
Mean hop delay	0.01s
Flow rate	Markov Model
Packet Size	64 bytes
Core router capacity	100Mbps
GW Capacity	2 Mbps
Transmission range of MR & GW	250m
Carrier sensing range	550m

Initially, a formulation is derived to compute the throughput from the simulation results. Thereafter this formulation is applied to obtain the throughput in the following section. First, a simple WMN is simulated without load management or GW scheduling. The throughput of this WMN is recorded. Thereafter the GW scheduling is performed, and the throughput is computed again. The performance of a conventional WMN is compared with a WMN with GW scheduling on basis of throughput, packet delay, and packet dropped parameters.

VII. ANALYSIS OF THE THROUGHPUT OF A WMN WITH THE PROPOSED LMS

This section investigates the throughput of WMN after applying each step of the proposed LMS. To compare the performance, it is necessary to create two WMN models namely Conventional WMN without LMS and WMN with proposed LMS. It can be observed that both these models can be derived from the system-level simulation model in Fig. 20 by making slight modifications in the IMW block.

A. Creation of a Simulation Model for Conventional WMN without LMS

To create a model of such a WMN, the simulation model of Fig. 20 is modified slightly. The IMW block as well as the Global APR table as explained in previous sections is disabled. The MRs are assigned to the GWs randomly and the respective APR tables of GWs are created accordingly. Thereafter, there is neither load monitoring nor load adjustments throughout the simulation period. This is the closest approximation to the conventional model of WMN without LMS.

B. Creation of Simulation Model of Load Management with Load Sharing

To obtain this model, the IMW block is modified such that all the blocks are enabled, and load monitoring is performed. Whenever the steady-state load condition is violated, the WMN performs load-sharing as explained in the IMW earlier. Since one of the major comparison parameters is the system throughput performance, the next subsection explains how throughput is calculated from the system utilization graph obtained through the simulation.

C. Process of Throughput Computation from the System Utilization Graph

This section explains how the throughput is calculated using the system utilization graph obtained from the MATLAB simulation. The system utilization graph is a part of the results displayed by the display block of the System-level simulation model in Fig. 20. It represents the average channel utilization of the entire WMN. The throughput computation is based on the following system parameters.

Core router bandwidth = 100 Mbps.

GW bandwidth = 2 Mbps.

Packet size=512 bits.

The throughput is determined using the relation.

$$\text{Throughput} = \text{Bandwidth} * \text{Utilization}.$$

D. Throughput Analysis with Load Sharing

This section presents the analysis of a WMN having the capability of load sharing incorporated in it. For uniformity and ease of comparison, the same scenarios of the previous section are reconsidered. For the simulation, the maximum capacity of GWs is assumed to be 2 Mbps. The first simulation on a WMN with load-sharing features pertains to the studies on the variation of throughput as a function of the number of MRs. As can be seen from the results of Fig. 21, initially the throughput improves with the increase in the number of MRs. The improved throughput is due to the fair scheduling of WMN. Regarding the throughput performance, it is interesting to note that the performance profiles exhibit a steep rise and slow decay characteristics.

It is worth mentioning here that a very similar trend was observed in [24] when they performed a similar study on fifth-generation cellular networks. It was observed by them that increasing the number of devices resulted in a decrease in data rates. They concluded that until better technology is devised, this decrease is bound to continue. This indicates that when a WMN gets congested and when all its GWs are utilized to their full potential, the throughput shall begin to drop. The relationship between the capacity of a WMN and the capacity of GW can be written as:

$$\text{The capacity of a WMN} = (\text{Capacity of GWs}) \times (\text{Number of GWs}).$$

E. Comparison of a Conventional WMN and a WMN with Proposed LMS

This section compares the throughput obtained in a conventional WMN without LMS with the throughput obtained in a WMN with the proposed LMS feature. The simulations have been performed keeping the total number of MRs fixed to 100 in each of the WMN scenarios but the number of GWs has been varied. This helps to study the effect of increasing the number of GWs on the throughput of WMN, keeping the number of MRs constant.

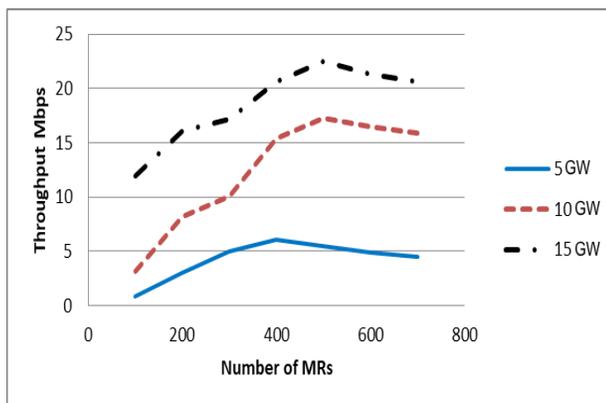


Fig. 21. Throughput of WMN with a Fixed Number of GWs and a Varying Number of MRs.

The results in Fig. 22 depict a continuous increase in the throughput after the application of the proposed LMS. The percentage increase in throughput of a WMN with 5GW 100MR after load sharing is 869%. In the case of a WMN with 10 GW 100 MR, the percentage increase is 893% after load sharing. For a WMN with 15GW 100 MR, the percentage increase in throughput is 1000% after load sharing!!!

The only difference in the simulation scenario between the Fig. 22 and Fig. 23 is the change in the total number of MRs to 200. Additionally, Fig. 23 also compares the throughput of a WMN obtained with a total number of 5 GW and 100 MRs.

The results of Fig. 23 reveal that a WMN with the same number of GWs but with a relatively larger total number of MRs exhibits better throughput performance. The throughput improvement of a WMN with 15GW 200MR is 1036% after load sharing. For a WMN with 10GW 200MR, the throughput improvement is 811% after load sharing. For a WMN with 15GW 200MR, the throughput improvement is 102.7% after load sharing. It can be noticed that if the number of MRs is fixed and only the number of GWs is increased, the gain in throughput is not as significant after performing load sharing. This is because when the number of GWs is increased but the number of MRs is fixed, then the GSS of every GW gets a lesser number of MRs. This results in a smaller number of MRs per GW and thus implying a relatively lesser gain in throughput.

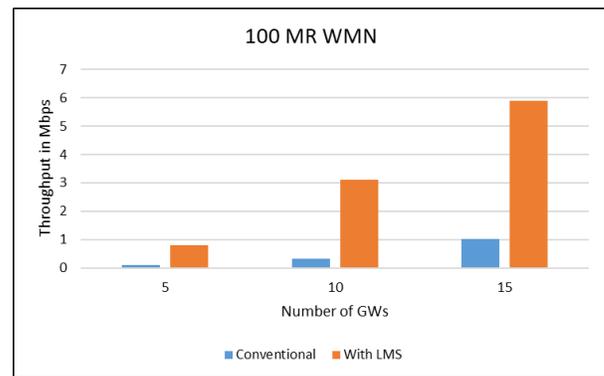


Fig. 22. Throughput Improvement of 100 MR WMN with Varying Number of GWs.

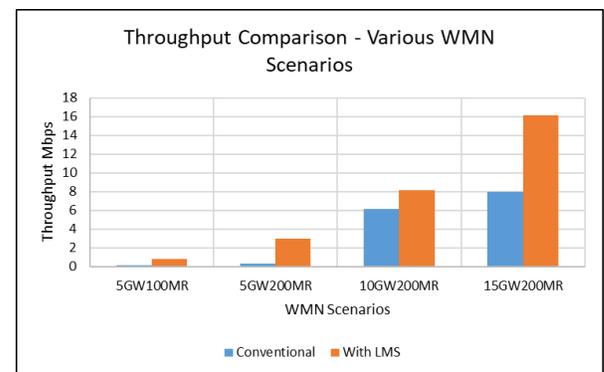


Fig. 23. Throughput Performance of LMS for Different WMN Scenarios.

The next section compares the parameter of packet drop before and after applying load sharing.

F. Analysis of Packet Loss

Since the proposed LMS relieves the congestion of GWs, it results in a reduced packet loss thereby leading to overall improved performance of a WMN. For the simulation, a WMN of 100 MRs is considered. Fig. 24 presents a comparative summary of the results obtained for the packet loss parameter. For the simulation results showed in Fig. 24, the total number of MRs remains constant at 100 while the number of GWs is varied from 5 to 15. The results of Fig. 24 depict an average 40% reduction in the packets dropped after applying the proposed LMS. This confirms the progressive improvement attributed to each constituent process of the proposed LMS thereby demonstrating the best performance.

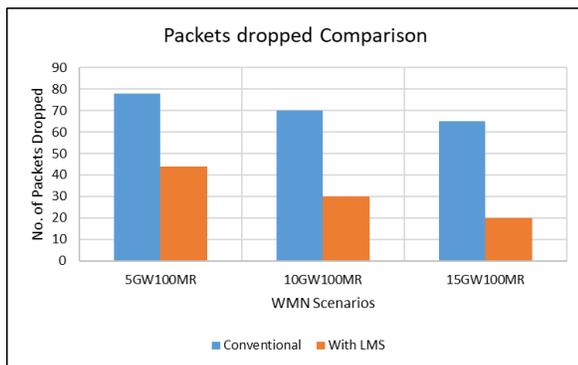


Fig. 24. Comparison of Packets Dropped With and Without LMS.

VIII. CONCLUSIONS AND FUTURE WORK

This paper has demonstrated the step-by-step process in the development of a simulation model of WMN using Simulink. The proposed simulation model was used to analyze the performance of the LMS proposed in this paper. It was found that there was an average 600% increase in the throughput of WMN after applying the LMS. The model also depicts a reduction of 40% in the number of packets dropped. The model is designed to facilitate the flexibility to increase or decrease the number of GWs within the mesh to support the increase and decrease in the number of hops for a particular router. The presented model will be helpful for researchers to analyze proposed techniques that involve many variations in the topology of WMN. The proposed model is also useful for the simulation of scheduling of Gateway and packet within a multi-hop multi gateway wireless network. This paper discusses the MATLAB model of conventional WMN and the implementation of the Load Management Scheme (LMS) on it. Using the proposed model, this paper has presented a comparative analysis of the throughput performance of WMN with and without the LMS. The uniqueness of the model presented in this paper is that it is scalable and flexible in terms of network topology parameters.

REFERENCES

[1] IEEE Standards Organisation, "IEEE Standard for Information Technology--Telecommunications and information exchange between systems--Local and metropolitan area networks--Specific requirements Part 11: IEEE Std. 802.11s-2011," Released on Sept. 10, 2011.

[2] Zigbee Alliance, Zigbee 3.0 standard specification. Released by Connectivity Standards Alliance, <https://zigbeealliance.org/wp-content/uploads/2019/11/docs-05-3474-21-0csg-zigbee-specification.pdf>, Release date: Nov 2019.

[3] K. N. Kapadia and D. D. Ambawade, "Congestion aware load balancing for multi-radio Wireless Mesh Network," International Conference on Communication, Information & Computing Technology (ICCICT), 2015, pp. 1-6, doi: 10.1109/ICCICT.2015.7045750.

[4] Son Pa, Satria Mandala, and Adiwijaya, "A new method for congestion avoidance in wireless mesh networks," The 2nd International Conference on Data and Information Science, Journal of Physics: Conference Series, Volume 1192, March 2019, pp 1-13, doi: 10.1088/1742-6596/1192/1/012062.

[5] Issariyakul, T., Hossain, E., "Introduction to Network Simulator 2 (NS2)", In Introduction to Network Simulator NS2, pp 21-41, Springer, Boston, MA, Oct 2011, https://doi.org/10.1007/978-1-4614-1406-3_2.

[6] Onur Özgün, Yaman Barlas, "Discrete vs. Continuous Simulation: When Does It Matter?," Proceedings of the 27th International Conference of The System Dynamics Society, July 26 – 30, pp 1-22, 2009.

[7] Sharma, R., Vashisht, V. and Singh, U., "Modelling and simulation frameworks for wireless sensor networks: a comparative study," IET Wirel. Sens. Syst., Vol 10: pp 181-197. Oct 2020.

[8] Mohammed Humayun Kabir, Syful Islam, Md. Javed Hossain, Sazzad Hossain, "Detail Comparison of Network Simulators," International Journal of Scientific & Engineering Research, Volume 5, Issue 10, pp 203, ISSN 2229-5518, October-2014.

[9] Qutaiba I. Ali, "Simulation Framework of Wireless Sensor Network (WSN) Using MATLAB/SIMULINK Software," in MATLAB - A Fundamental Tool for Scientific Computing and Engineering Applications - Volume 2. London, United Kingdom: IntechOpen, September 26th, 2012.

[10] Alasaad, A., Gopalkrishnan, S., Leung, V.C.M., "Mitigating Load Imbalance in Wireless Mesh Networks with Mixed Applications Traffic Types," Proceedings of the Global Telecommunications Conference, IEEE Vol. no. 1-5, pp. 6-10, December 2010.

[11] Ernst, Jason B., Denko, Miso K. "Fair Scheduling with Multiple Gateways in Wireless Mesh Networks," IEEE International Conference on Advanced Information Networking and Applications, AINA, pp.106-112, 2009.

[12] Bruno, R., Conti, M., and Pinizzotto, A. "Routing Internet Traffic in Heterogeneous Mesh Networks: Analysis and Algorithms," Elsevier Journal of Performance Evaluation, vol. 68, no. 9, pp. 841-58, 2011.

[13] Jun, Jangeun and Sichertiu, M.L. "The Nominal Capacity of Wireless Mesh Networks," IEEE Journal of Wireless Communications, vol. 10, no. 5, pp. 8-14, Oct 2003.

[14] Das, Saumitra & Pucha, Himabindu, and Hu, Y. "Mitigating the Gateway Bottleneck via Transparent Cooperative Caching in Wireless Mesh Networks," ACM Journal of Ad Hoc Networks, Volume 5, Issue 6, August 2007 pp 680-703, <https://doi.org/10.1016/j.adhoc.2006.11.004>.

[15] A. Ouni, H. Rivano, and F. Valois, "Capacity of Wireless Mesh Networks: Determining Elements and Insensible Properties," IEEE Wireless Communication and Networking Conference Workshops, 2010, pp. 1-6, doi: 10.1109/WCNCW.2010.5487652.

[16] Ken Kutzler, "Building a Core Router for the Next Decade," Nokia Blog, 22 May 2012, available at: <https://www.nokia.com/blog/building-core-router-next-decade/>.

[17] M Kiran Sastry, Arshad Ahmad Khan Mohammad, and Arif Mohammad Abdul, "Optimized Energy-efficient Load Balance Routing Protocol for Wireless Mesh Networks" International Journal of Advanced Computer Science and Applications (IJACSA), Vol 12, issue 8, pp 605-610, 2021.

[18] Chai Y, Zeng XJ. The development of green wireless mesh network: A survey. Journal of Smart Environments and Green Computing, Vol 1, pp 47-59, 2021. <http://dx.doi.org/10.20517/jsegc.2020.05>.

[19] J. R. Parvin, "An Overview of Wireless Mesh Networks", in Wireless Mesh Networks - Security, Architectures, and Protocols. London, United Kingdom: IntechOpen, 2019 [Online]. Available: <https://www.intechopen.com/chapters/66938>.

- [20] Zhou, P., "On optimizing wireless mesh networks: from theoretical capacity analysis to practical algorithm design," Ph.D. thesis submitted at UC San Diego. ProQuest ID: umi-ucsd-1898. 2008. Available at: <https://escholarship.org/uc/item/9s90t2hq#main>.
- [21] IEEE Standards Association, "Institute of Electrical and Electronics Engineers (2007) IEEE 802.11-2007: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications," Released in June 2007.
- [22] Mathworks, "Simulink: Simulation and model-based design," 2018, https://in.mathworks.com/help/simulink/index.html?s_tid=CRUX_lftnav
- [23] UC Berkley, USC/ISI, LBL, and Xerox PARC "The ns manual," ed. by Kevin Fall and Kannan Varadhan, Chapter 16, page 152, 4 November 2011.
- [24] Boccardi F., Heath R.W., Lozano, A., Marzetta T.L., Popovski P. "Five Disruptive Technology Directions for 5G," Communications Magazine, IEEE, vol. 52, no. 2, pp. 74,80, February 2014.

A Model for Classification and Diagnosis of Skin Disease using Machine Learning and Image Processing Techniques

Shaden Abdulaziz AlDera¹
Department of Computer Science
College of Computer
Qassim University, Buraydah 51452
Saudi Arabia

Mohamed Tahar Ben Othman² 
BIND Research Group, IEEE Senior Member
Department of Computer Science
College of Computer, Qassim University
Buraydah 51452, Saudi Arabia

Abstract—Skin diseases are a global health problem that is difficult to diagnose sometimes due to the disease's complexity, and the time-consuming effort. In addition to the fact that skin diseases affect human health, it also affects the psycho-social life if not diagnosed and controlled early. The enhancement of images processing techniques and machine learning leads to an effective and fast diagnosis that help detect the skin disease early. This paper presents a model that takes an image of the skin affected by a disease and diagnose acne, cherry angioma, melanoma, and psoriasis. The proposed model is composed of five steps, i.e., image acquisition, preprocessing, segmentation, feature extraction, and classification. In addition to using the machine learning algorithms for evaluating the model, i.e., Support Vector Machine (SVM), Random Forest (RF), and K-Nearest Neighbor (K-NN) classifiers, and achieved 90.7%, 84.2%, and 67.1%, respectively. Also, the SVM classifier result of the proposed model was compared with other papers, and mostly the proposed model's result is better. In contrast, one paper achieved an accuracy of 100%.

Keywords—Skin disease; image processing; classification; machine learning; diagnosis; SVM; RF; K-NN; acne; cherry angioma; melanoma; psoriasis

I. INTRODUCTION

Nowadays, imaging is used in medical science extensively, so before any surgery or treatment decision a preliminary knowledge can be determined, and diagnosis can be done. For this, imaging in medicine has become a tool to start most of the disease treatment cycle, starting from detection passing through evaluation, and ending with the treatment decision. Skin disease is one of these medical areas where images play a role in detecting, diagnosing, and treating the disease [1]. In recent years, skin diseases have increased and begun to be a global health problem [2]. Those who suffer from skin diseases without disease diagnosis may diminish their life quality and have a negatively psycho-social impact [3].

In fact, skin diseases are difficult to diagnose due to the complexity of human skin. Also, the lack of expertise may lead to misdiagnosis or overdue diagnoses. Diagnosis of skin diseases at the health center may take a long time and require domain expertise, which causes physical and financial costs. On the other side, machine learning and image processing

techniques can help achieve high accuracy in skin diagnosing at the initial stage. Images processing plays an effective role in diagnosis the skin diseases with the help of libraries such as OpenCV, Scikit-Image, and NumPy. Afterward, machine learning algorithms such as SVM, RF, and K-NN are used for the classification task. Combining these techniques will save time and reach a quicker and more trusted diagnostic than typical procedures like patch tests and biopsy [4]. Due to the limitation of the existing models that diagnose different skin diseases, this proposed model studied four skin diseases. This research work aims to build a model that provides an easy, fast and efficient solution for skin disease diagnosis, i.e., acne, cherry angioma, melanoma, and psoriasis, using image processing and machine learning techniques.

The building of the proposed model passes through multi-steps of image processing including, acquiring images, preprocessing images, i.e., resizing images, color transformation, de-noising, and normalization, segmentation, and feature extraction. In the end, train the model with traditional machine learning algorithms, e.g., SVM, RF, and K-NN. Several papers conducted on this paper's topic had been focused on the use of machine learning and image processing to classify skin cancer. Thus, this paper proposed a model to diagnose other common diseases in addition to skin cancer. Furthermore, the proposed model tested cherry angioma disease, which is very rarely tested in the previous research.

This paper is organized into the following sections: Section 2 reviews the previous work, Section 3 presents the methodology, Section 4 shows the obtained results, and Section 5 is the conclusion of this work.

II. PREVIOUS WORK

Many researchers have proposed a model that combines image processing and machine learning algorithms techniques to classify and diagnose several skin diseases.

Hameed, Shabut, and Hossain [5] implemented a system that classifies healthy, acne, eczema, psoriasis, benign, and melanoma (malignant) skin diseases. The system was built based on image processing techniques. To enhance images, the authors used an algorithm called Dull Razor to remove hair from skin images and then applied the Gaussian filter to

smooth it. After that, in the segmentation task, they firstly discarded any non-skin area, then applied Otsu's thresholding on the area of skin in order to segment the disease lesion. Moreover, they used color spaces like red, green, and blue (RGB), hue, saturation, and value (HSV), luma, blue, and red (YCbCr) which means separate brightness i.e. luma from color, and grayscale to extract color features. Also, they used neighborhood gray-tone difference matrix (NGTDM) and Gray Level Co-occurrence Matrix (GLCM) techniques to extract the texture features. In the end, they classified the diseases using SVM and obtained an accuracy of 83%.

Additionally, Sinthura S. et al. [6] propose a method to detect skin diseases. Their method indicates using the adaptive filter to remove the noises, then converting it to grayscale color. Besides, they used Otsu's thresholding technique to segment the disease lesion. Furthermore, they used GLCM to extract the texture features. Finally, to validate their proposal, they used the SVM classifier and achieved an accuracy of 89%. In image classification-based color, the researchers in [7] train a model for detecting and classifying various skin diseases using the K-NN classifier. They use color models to extract features, including the HSV and the lightness, red/green, and blue/yellow (L^*a^*b) color models. Their results showed that the HSV color model is more efficient with 91.80% accuracy than the L^*a^*b color model with 81.60% accuracy. Moreover, Ahmed, Ema, and Islam [8] propose a new automated system using the Transductive SVM (TSVM) to classify 24 types of skin disease. The proposed system uses a hybrid genetic algorithm to segment the image. Also, they used ant colony optimization (ACO-GA) and GLCM to extract its features. Their work achieved 95% accuracy.

A method was carried out to apply pre-trained Convolutional Neural Networks (CNN) to extract features for skin diseases. The paper by ALenezi [9] proposes a system using the pre-trained CNN AlexNet to extract skin disease features and SVM to classify the diseases. The system was built with a dataset of 80 images of these skin types; melanoma, psoriasis, eczema, and healthy skin. Her system was tested on 20 images and achieved an accuracy of 100%. The same method used by researchers [10] present an intelligent expert system for classifying 9144 skin lesions, i.e., acne, eczema, benign or malignant (melanoma), and healthy skin images. For extracting the lesion's features, they used a pre-trained CNN model AlexNet. Their system result achieved an accuracy of 86.21% by using the SVM classifier, where divided dataset in the ratio 70:30 for train and test set.

Another study by Hajgude et al. [11] proposes a solution to detect 408 images of eczema, impetigo, melanoma skin diseases, and a class named other images. They build the model using these techniques: median filter to remove the noises, the Otsu method to segment lesions, 2D Wavelet transform to extract features like entropy and standard deviation, and GLCM to extract texture features like contrast and correlation. They used SVM and CNN classifiers to classify the diseases and obtained an accuracy of 90.7% and 99.1%, respectively. Authors in [12] describe the CNN classifier and the major libraries for image processing. Then, they use CNN to classify and diagnose skin disease with an accuracy of 70%, even though using a large dataset may

increase accuracy by more than 90%. Authors in [13] propose a web system to diagnose skin diseases in Ghana (Africa). Their study includes a CNN to classify 254 images of diseases, i.e., atopic dermatitis, acne vulgaris, and scabies. In the end, they reached an accuracy for each disease of 88%, 85%, and 84.7%, respectively. The proposed system just takes 0.0001 seconds to diagnose, which may expedite more patients' diagnoses than a diagnosis in the clinic.

Most of the previous studies prove the efficacy of using SVM and CNN. Furthermore, the studies evinced the image processing plays a key role in helping to classify various skin diseases. Moreover, increasing the number of images may positively affect the classification due to the increased training model.

III. METHODOLOGY

This paper demonstrates the classification of several types of skin disease to diagnose the lesions such as acne, cherry angioma, melanoma, and psoriasis. Accordingly, the processes involved in identifying these skin diseases are preprocessing, segmentation, feature extraction, and classification. The following points show the datasets used and explain the proposed and techniques of this work.

A. Dataset

Due to the privacy of medical records, collecting images is a challenging task. Therefore, the images gathered from available resources: the dermnet NZ [14] and atlas dermatologico [15]. In this work, the dataset consists of 377 images of four different disease classes: acne, cherry angioma, melanoma, and psoriasis. Fig. 1 shows a sample of each class. Table I lists the number of images of each class.

1) *Diseases definition*: In the following, a brief definition of each disease studied in this work, as mentioned in the dermnet NZ website [14].

a) *Acne*: It is a common chronic disorder, often confined to the face, but it may happen in the chest, back, and neck. Acne may occur in children and adults of all ages. However, acne is caused due to a combination of several factors such as familial tendency, acne bacteria, and hormones. Acne could be characterized as blackheads and whiteheads.

b) *Cherry angioma*: The reason of cherry angioma is unknown. It is very prevalent in males and females of any age, while it markedly increases in people from about the age of 40. However, cherry angioma may be in red or purple, or blue color. Also, it could be scattered overall body surface parts.

c) *Melanoma*: It is the gravest form of skin cancer. It happened due to the uncontrolled growth of melanocytes (pigment cells). Melanoma may occur at any age but is very rare in children. However, the features of melanoma could be having several colors like blue, brown, red, etc.

d) *Psoriasis*: It is a chronic inflammation of a skin condition. It affects males and females at any age. It is characterized by symmetrically distributed, red color, scaly plaques with well-defined edges.



Fig. 1. Dataset Samples

TABLE I. DATASET DISTRIBUTION

Disease	Dataset	Total
Acne	Dermnet NZ dataset, Atlas dermatologico	80
Cherry angioma	Dermnet NZ dataset	37
Melanoma	Dermnet NZ dataset, Atlas dermatologico	80
Psoriasis	Dermnet NZ dataset, Atlas dermatologico	180
Total		377

B. Proposed Methodology

This section presents the processes of this proposed model and the techniques used. The architecture of this model is shown in Fig. 2. The procedure of the proposed model is described in the following points:

- Import the train set images and process it through: preprocessing the images, then segmenting the lesion from the remaining normal skin, and after that extracting its features.
- Import the test set images and process them in the same way as step 1.
- The features that extract from the train set images are stored in a knowledge base.
- Compare the features that extract from the test set image with the feature stored in the knowledge base.
- Classify images using machine learning algorithms i.e., SVM, RF, K-NN.
- Diagnosis of the disease.
- When the user uploads an image, it is will pass through the same processing of the test set image.

1) *Image processing*: Image processing is a technique that manipulates and analyzes an image received from a camera or sensors. Therefore, the image processing's main objective is to enhance the image's quality and extract its information in order to be more interpretable by a human or machine perception. Nowadays, many image processing techniques are incorporated as they turn out to be strong computational methodologies and strong potential to be effective in healthcare and all fields [16]. In the following, describe the image processing techniques used in this work.

a) *Preprocessing*: The first step in processing the skin disease images is preprocessing in order to enhance images. In this work, firstly, all the images were resized to 250X250.

After that, due to the noises in skin images, a de-noising technique called a median filter was applied. The median filter is the most filter the researcher used according to the advantage of preserving the edges of the image [17]. Further, the color images converted to a grayscale color model for segmentation and feature extraction tasks. Finally, the images pixels' values normalized between 1 and 0. In Fig. 3, column A shows the original images with a size of 250X250. Next, column B shows the images after applying the median filter. Also, column C shows images converted to grayscale.

b) *Segmentation*: The segmentation task is a process of partitioning the lesion region from the skin. This segment gains based on similarity or difference of pixels properties like color, sharpness, brightness, or intensity of an image [18][19]. Based on this work, it's a challenging task due to the several diseases the proposed test. Also, the key problem is the entire lesion's color may be similar, and the lesion's boundaries may be fuzzy, besides the complexity of the skin itself. To address this problem, the Otsu's thresholding is used to create a mask (binary image), then applied it to the grayscale image. Some of the results are shown in Fig. 3. Column D shows the binary masks, and column E shows the final segmentation results.

- Otsu's: Otsu's is the most popular threshold segmentation technique, and it is applied to a grayscale image. Unlike the manual threshold, it automatically compares the minimize weighted within gray classes variance to find the optimal threshold value. Since the threshold value was determined, the lesion can be segmented from the normal skin region [20][21].

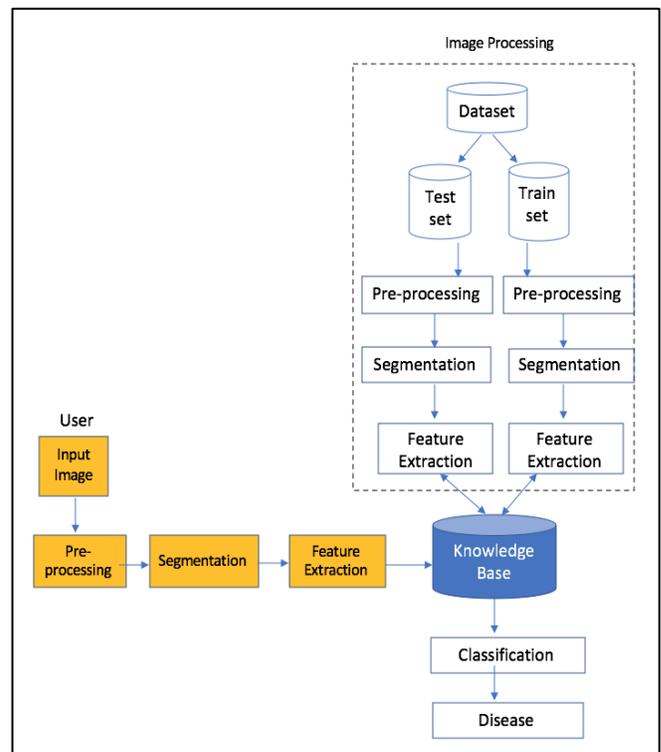


Fig. 2. Model Architecture.

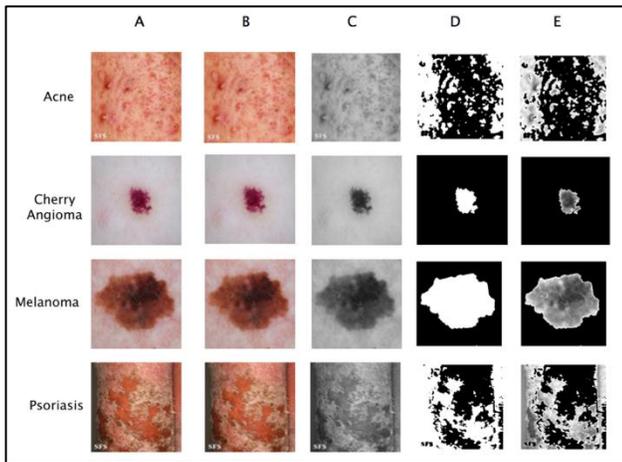


Fig. 3. Samples Results of Images Processing Techniques.

c) *Feature Extraction*: Feature extraction is a technique that plays a significant role in image processing. It is used the image or the segmented lesion to extract the characteristics' features that represent the information of that image for classification tasks. However, texture feature is a type of feature that can be used in recognition of an image by describing the visual image's surface. Textures are complex patterns composed of many characteristics, including size, color, brightness, slope, etc. [22]. In this work, feature extraction is performed using Gabor and Entropy techniques to extract texture features and the Sobel technique for edge

features. All the feature extraction techniques' parameters used in this proposed model are listed in Table II.

- **Gabor**: Gabor filter is a linear filter used to extract texture features. It is the most commonly used in image processing and image texture analysis [23]. Essentially, Gabor is a band-pass that extracts patterns in a signal at specific frequencies [24].
- **Entropy**: It measures the expectation of the quantity of information in the grayscale image. It calculates the center pixel of all the neighboring pixels in the kernel's window [25].
- **Sobel**: Sobel is a filter used to detect the image's edge features. It has the power on processing with more minor time consumption, less in loss of edge, and strongly resists the noise [26]. In this work, the Sobel is applied that utilized two kernels to obtain the horizontal(Gx) and vertical(Gy) approximation of the derivation of the grayscale skin diseases images. These kernels were convolved to overlay the image's pixels.

2) *Classification*: The classification task classifies data into distinct known classes using machine learning algorithms to predict the disease. Once features are extracted, it is given as input to the classifier model. When the model is accurate, it is used to classify new images that are a member of the trained disease classes. The proposed model used the traditional well-known classifiers to conduct the experiments, i.e., SVM, RF, and K-NN.

TABLE II. THE PARAMETERS OF FEATURE EXTRACTION TECHNIQUES

Feature Extraction Technique	Formula	Parameters	Value
Gabor	$G(x, y, \lambda, \theta, \phi, \sigma, \gamma) = \exp\left(-\frac{(x^2 + \gamma^2 y^2)}{2\sigma^2}\right) \cos(2\pi \frac{x}{\lambda} + \phi)$, where $x = x \cos \theta + y \sin \theta$ $y = -x \sin \theta + y \cos \theta$	λ : wavelength of the cosine multiplier. θ : Frequency of alternations in degrees. ϕ : Phase offset of the sinusoidal function. σ : Sigma / standard deviation of the Gaussian envelope. γ : Spatial aspect ratio and specifies the ellipticity of the support of Gabor function.	$\pi/4 = 3.14/4$ $4 * \pi = 4 * 3.14$ 0 1 and 3 0.5
Entropy	$H(X) = - \sum_{i=1}^n P(x_i) \log_b P(x_i)$, where $\log_b = 2$	Kernel	3X3
Sobel	$G_{x,y} = \sqrt{G_x^2 + G_y^2}$	Two Kernels: Gx and Gy	3X3 and 3X3

IV. EXPERIMENTAL RESULTS

This section shows the evaluated experiments to measure the performance of this model. The proposed model's experiments were implemented using the python language in Spyder environment from anaconda, along with several libraries, i.e., Scikit-learn to perform machine learning algorithms, OpenCV, Glob, and Os to perform image processing, Skimage to perform filters, and Matplot for the visualization.

A. Evaluation Measure

The model performance was analysis with several measures. These measures are formulated as in (1), (2), (3), and (4):

$$\text{Accuracy} = (TP + TN) / (TP+FP+TN+FN) \quad (1)$$

$$\text{Precision} = TP / (TP+FP) \quad (2)$$

$$\text{Recall} = TP / (TP+FN) \quad (3)$$

$$\text{F1-score} = 2 * (\text{precision} * \text{recall} / (\text{precision} + \text{recall})) \quad (4)$$

Where TP is true positive, TN is a true negative, FP is false positive, and FN is a false negative.

B. Splitting Dataset and Classifiers Parameters Results

To detect the behavior of machine learning algorithms, it needs to test the model on data that is not used in the training process. Toward this, the dataset splits into two sets: training set and testing set. Table III lists three experiments of splitting and lists the accuracy results of each classifier.

TABLE III. THE SPLITTING DATASET WITH TRAIN AND TEST RESULTS

Splitting	#images	Classifiers	Parameters	Training	Testing
Train: 70% Test: 30%		SVM	Kernal = 'rbf' n_estimators=50 n_neighbors=5	92.3%	78%
	Train: 263	RF		100%	73.6%
	Test: 114	K-NN		73.7%	67.5%
Train: 75% Test: 25%	Train: 282	SVM	Kernal = 'rbf' n_estimators=50 n_neighbors=5	91.4%	77.8%
	Test: 95	RF		100%	74.7%
		K-NN		72.6	68%
Train: 80% Test: 20%		SVM	Kernal = 'rbf' n_estimators=50 n_neighbors=5	91%	90.7%
	Train: 301	RF		100%	84.2%
	Test: 76	K-NN		73%	67.1%

TABLE IV. CLASSIFIERS PARAMETERS AND THEIR RESULTS

Classifiers	Parameters	Values	Accuracy
SVM	Kernel Regularization(C) Degree	'linear', 'rbf', 'poly', 'sigmoid' 1.0 3	'Linear': 81.5% 'rbf': 90.7% 'poly': 71% 'sigmoid': 63%
RF	n_estimators	30,40,50,60	30: 73.6% 40:80% 50: 84.2% 60: 82.8%
K-NN	n_neighbors	3,4,5	3: 64% 4: 63% 5: 67.1%

Each ML classifier has several possible values of each corresponding parameter that value may affect on accuracy result. However, the value that obtains the highest accuracy result will be used in the proposed model. The possible parameters and their values are illustrated with the accuracy results in Table IV. All these experiments were tested on the dataset that split into 80% for training and 20% for testing.

C. Classifiers Performance Matrix

This section shows the confusion matrix of each classifier. It is represented in a table style to describe the model's classifier performance. However, it presents the prediction results on the test set data. The matrix size is based on the number of classes; in this case, the matrix size is 4X4. The rows indicate the true class, while the columns indicate the prediction for each class. While The diagonal of the matrix points to the number of images that are correctly classified. Fig. 4 displays the confusion matrices of the classifiers of SVM, RF, and K-NN as shown in A, B, and C, respectively. For example, in SVM, a total of 16 acne images, it classified 14 images correctly and 2 images were misclassified as psoriasis.

From these confusion matrices, could measure the accuracy, precision, recall, and F1-score of each disease with each classifier as shown in Table V. Among the three classifiers, SVM is superior performance on the accuracy rate of cherry angioma, melanoma, and psoriasis. While acne obtained the highest accuracy rate using the RF classifier. In contrast, K-NN has the worst results accuracy in all these diseases.

D. Classification Experiment

The proposed model was validated using SVM, RF, and K-NN classifiers with the evaluation measures, i.e., accuracy, precision, recall, and f1-score. Table VI lists all measurement results of each classifier. It is observed that the SVM has superiority in classifying skin diseases over others in accuracy, precision, recall, and f1-score of 90.7%, 91%, 90.8%, and 90.8%, respectively. At the same time, the K-NN has obtained the worst results.

E. Comparison Results

To our knowledge, there are no study experiments on the same diseases of this work. Also, due to different and unavailable datasets, the proposed model was compared with the research that tested some of the diseases studied in this work. Table VII details the comparison research [5], [6], [9], and [10] toward their studied diseases and the image processing techniques they used.

Basically, many techniques are available for image processing and classification tasks, and it used in several skin diseases classification research. Among that, the comparison research in Table VII are probably the most papers close to the techniques used and to the diseases tested in this work.

In the comparison papers, they present in two ways of image processing: manual and automatic. In manual image processing, researchers [5], [6] follow the same processing steps with different preprocessing filters techniques; Dull Razor with Gaussian in [5], and Median in [6]. Further, they used the same segmentation technique, i.e., Otsu's. Another essential point, studies have shown that extraction techniques play a key role in extracting the appropriate features for the

classification task. For that, [5] extract the texture features using GLCM and NGTDM, and extract color features using color spaces. Similarly, researchers [6] extract texture features using GLCM. On the other hand, both researchers [9] and [10] utilized the automatic image processing by CNN with a pre-train AlexNet.

Therefore, Table VIII lists the SVM classifier accuracy of this proposed model with the comparison research model's accuracy that used the same classifier, i.e., SVM, and different image processing methods. All these papers showed a promising high accuracy rate in diagnosing diseases above

83%. However, it was observed the proposed model is higher than [5], [6], and [10] with 90.7% accuracy.

Thus, probably the proposed model's performance is better since this proposed model used multi techniques to extract a combination of features, i.e., Gabor and Entropy for texture features and Sobel for edge features. In contrast, the paper [9] is superior to the proposed model with 100% accuracy. Despite the fact, that they trained the model on 80 images, while the proposed model was trained on 301 images. Also, the proposed model tested cherry angioma disease, whereas the other studies did not test it.

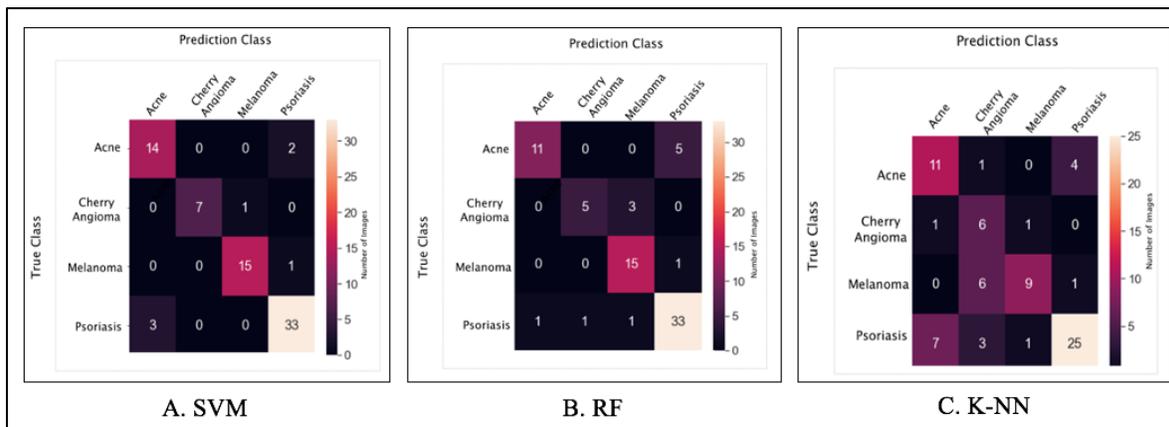


Fig. 4. The Confusion Matrices OF SVM, RF, AND K-NN Classifiers.

TABLE V. DISEASE CLASSIFICATION RATE

		Accuracy	Precision	Recall	F1-score
SVM	Acne	82%	82%	88%	85%
	Cherry Angioma	100%	100%	88%	93%
	Melanoma	94%	94%	94%	94%
	Psoriasis	92%	92%	92%	92%
RF	Acne	92%	92%	69%	79%
	Cherry Angioma	83%	83%	62%	71%
	Melanoma	79%	79%	94%	86%
	Psoriasis	85%	85%	92%	88%
K-NN	Acne	58%	58%	75%	60%
	Cherry Angioma	38%	38%	75%	52%
	Melanoma	82%	82%	56%	67%
	Psoriasis	83%	83%	61%	71%

TABLE VI. CLASSIFIERS RESULTS

Classifier	Accuracy	Precision	Recall	F1-Score
SVM	90.7%	91%	90.8%	90.8%
RF	84.2%	84.8%	84.2%	83.8%
K-NN	67.1%	72.8%	67.1%	68.4%

TABLE VII. THE COMPARISON RESEARCH

Ref.	Disease	Segmentation technique	Feature techniques
Hameed N et al. [5]	Acne, Psoriasis, Melanoma, Benign, Eczema, and Healthy Skin.	Otsu's	GLCM, NGTDM, and color spaces
Sinthura S. et al. [6]	Acne, Psoriasis, Melanoma, and Rosacea.	Otsu's	GLCM
ALenezi N. [9]	Psoriasis, Melanoma, Eczema, and Healthy Skin.	-	CNN: AlexNet.
Hameed N et al. [10]	Acne, Melanoma, Benign, Eczema, and Healthy Skin.	-	CNN: AlexNet.
Proposed Model	Acne, Cherry angioma, Melanoma, and Psoriasis.	Otsu's	Entropy, Gabor, and Sobel.

TABLE VIII. COMPARISON OF ACCURACY

Ref.	Accuracy
Hameed N et al. [5]	83%
Sinthura S. et al. [6]	89%
ALenezi N. [9]	100%.
Hameed N et al. [10]	86.21%
Proposed Model	90.7%

The main problem encountered in developing the proposed model was the few of availability of the image of the diseases tested. In addition, the few papers that study various types of skin diseases. May training more images increase the accuracy and make the model more accurate to diagnose new images, as well as selecting appropriate techniques to extract useful features.

V. CONCLUSION

This paper proposed a model that provides the classification of different types of skin diseases: acne, cherry angioma, melanoma, and psoriasis. According to the previous works in this area, there is a scarcity in the studies on these diseases as most research focuses on skin cancer. This model was conducted through image processing techniques and machine learning algorithms on a total of 377 images. The dataset is divided into 301 images for the train set and 76 images for the test set. Firstly, the processing techniques are applied to images in several steps: preprocessing including resizing images, removing the noise using the median filter and converting the image to grayscale, then separating the infected area using Otsu's, and extracting its features using Gabor, Entropy, and Sobel. Secondly, the model was evaluated using SVM, KNN, and RF classifiers in terms of accuracy, precision, recall, and f1-score. However, the proposed model results show that the SVM accomplished higher accuracy with 90.7% than RF and K-NN. At the same time, RF and K-NN achieved 84.2% and 67.1% accuracy, respectively. The result of the proposed model using the SVM classifier achieved better accuracy than the comparison research' accuracy. In contrast, one paper outperformed the proposed model's accuracy. Since it was not possible to collect a skin disease dataset locally, finding a public source with multi images of diseases was the biggest challenge during this work. This model can accomplish higher accuracy by using more dataset images. Moreover, programmers can utilize the model to develop a smartphone application to diagnose these skin diseases easily and early.

REFERENCES

[1] L. Bajaj, H. Kumar, and Y. Hasija, "Automated System for Prediction of Skin Disease using Image Processing and Machine Learning," *Int. J. Comput. Appl.*, vol. 180, no. 19, pp. 9–12, 2018, doi: 10.5120/ijca2018916428.

[2] T. Vos et al., "Global, regional, and national incidence, prevalence, and years lived with disability for 328 diseases and injuries for 195 countries, 1990–2016: A systematic analysis for the Global Burden of Disease Study 2016," *Lancet*, vol. 390, no. 10100, pp. 1211–1259, 2017, doi: 10.1016/S0140-6736(17)32154-2.

[3] E. Wojtyna, P. Łakuta, K. Marcinkiewicz, B. Bergler-Czop, and L. Brzezińska-Wcisło, "Gender, body image and social support: Biopsychosocial determinants of depression among patients with

psoriasis," *Acta Derm. Venereol.*, vol. 97, no. 1, pp. 91–97, 2017, doi: 10.2340/00015555-2483.

[4] V. Pugazhenthil et al., "Skin Disease Detection And Classification," *Int. J. Adv. Eng. Res. Sci.*, vol. 6, no. 5, pp. 396–400, 2019.

[5] N. Hameed, A. Shabut, and M. A. Hossain, "A Computer-Aided diagnosis system for classifying prominent skin lesions using machine learning," 2018 10th *Comput. Sci. Electron. Eng. Conf. CEEC 2018 - Proc.*, pp. 186–191, 2019, doi: 10.1109/CEEC.2018.8674183.

[6] S. S. Sinthura, K. R. Sharon, G. Bhavani, L. Mounika, and B. Joshika, "Advanced Skin Diseases Diagnosis Leveraging Image Processing," *Proc. Int. Conf. Electron. Sustain. Commun. Syst. ICESC 2020*, no. Icesc, pp. 440–444, 2020, doi: 10.1109/ICESC48915.2020.9155914.

[7] A.V.Ubale and P. L. Paikrao, "Detection and Classification of Skin Diseases using Different Color Phase Models," *Int. Res. J. Eng. Technol.* e-ISSN, vol. 06, no. 07, pp. 3658–3663, 2019.

[8] M. H. Ahmed, R. R. Ema, and T. Islam, "An Automated Dermatological Images Segmentation Based on a New Hybrid Intelligent ACO-GA Algorithm and Diseases Identification Using TSVM Classifier," 1st *Int. Conf. Adv. Sci. Eng. Robot. Technol. 2019, ICASERT 2019*, vol. 2019, no. Icasert, 2019, doi: 10.1109/ICASERT.2019.8934560.

[9] N. S. Alkolifi Alenezi, "A Method of Skin Disease Detection Using Image Processing and Machine Learning," in *Procedia Computer Science*, 2019, vol. 163, pp. 85–92, doi: 10.1016/j.procs.2019.12.090.

[10] N. Hameed, A. M. Shabut, and M. A. Hossain, "Multi-Class Skin Diseases Classification Using Deep Convolutional Neural Network and Support Vector Machine," *Int. Conf. Software, Knowl. Information, Ind. Manag. Appl. Ski.*, vol. 2018-Decem, 2019, doi: 10.1109/SKIMA.2018.8631525.

[11] M. J. Hajgude, A. Bhavsar, H. Achara, and N. Khubchandani, "Skin disease detection using Image Processing with data mining and deep learning," *IRJET*, vol. 6, no. 4, pp. 4363–4366, 2019.

[12] J. Rathod, V. Wazhmode, A. Sodha, and P. Bhavathankar, "Diagnosis of skin diseases using Convolutional Neural Networks," *Proc. 2nd Int. Conf. Electron. Commun. Aerosp. Technol. ICECA 2018*, no. Iceca, pp. 1048–1051, 2018, doi: 10.1109/ICECA.2018.8474593.

[13] S. Akyeramfo-Sam, A. Addo Philip, D. Yeboah, N. C. Nartey, and I. Kofi Nti, "A Web-Based Skin Disease Diagnosis Using Convolutional Neural Networks," *Int. J. Inf. Technol. Comput. Sci.*, vol. 11, no. 11, pp. 54–60, 2019, doi: 10.5815/ijitcs.2019.11.06.

[14] "DermNet NZ." website: dermnetnz.org, license: <https://creativecommons.org/licenses/by-nc-nd/3.0/nz/legalcode> (accessed Jun. 01, 2021).

[15] S. F. da Silva and D. B. Calheiros., "Dermatology Atlas." atlasdermatologico.com.br (accessed Jun. 01, 2021).

[16] S. Perumal and T. Velmurugan, "Preprocessing by Contrast Enhancement Techniques for Medical Images," *International Journal of Pure and Applied Mathematics*, vol. 118, no. 18, pp. 3681–3688, 2018.

[17] L. Kapoor and S. Thakur, "A Survey on Brain Tumor Detection Using Image Processing Techniques," 2017 7th *Int. Conf. Cloud Comput. Data Sci. Eng. - Conflu.*, pp. 582–585, 2017, doi: 10.1109/CONFLUE NCE.2017.7943218.

[18] M. M. Lone and S. Hussain, "A Survey on Image Segmentation Techniques," *Int. Res. J. Eng. Technol.*, vol. 05, 2018, doi: 10.1007/978-3-030-32150-5_112.

[19] H. Seo et al., "Machine learning techniques for biomedical image segmentation: An overview of technical aspects and introduction to state-of-art applications," *arXiv*, pp. 1–35, 2019.

[20] H. Sen and A. Agarwal, "A comparative analysis of entropy based segmentation with Otsu method for gray and color images," *Proc. Int. Conf. Electron. Commun. Aerosp. Technol. ICECA 2017*, vol. 2017-Janua, pp. 113–118, 2017, doi: 10.1109/ICECA.2017.8203655.

[21] Y. Wang, "Improved OTSU and adaptive genetic algorithm for infrared image segmentation," *Proc. 30th Chinese Control Decis. Conf. CCDC 2018*, pp. 5644–5648, 2018, doi: 10.1109/CCDC.2018.8408116.

[22] S. Kolkur and D. R. Kalbande, "Survey of texture based feature extraction for skin disease detection," *Proc. 2016 Int. Conf. ICT Business, Ind. Gov. ICTBIG 2016*, 2017, doi: 10.1109/ICTBIG.2016.7892649.

- [23] H. Ji, Z. Shen, and Y. Zhao, "Digital Gabor filters do generate MRA-based wavelet tight frames," *Appl. Comput. Harmon. Anal.*, vol. 47, no. 1, pp. 87–108, 2019, doi: 10.1016/j.acha.2017.08.001.
- [24] S. T. H. Rizvi, G. Cabodi, P. Gusmao, and G. Francini, "Gabor filter based image representation for object classification," *Int. Conf. Control. Decis. Inf. Technol. CoDIT 2016*, pp. 628–632, 2016, doi: 10.1109/CoDIT.2016.7593635.
- [25] F. Hrzić, I. Štajduhar, S. Tschauner, E. Sorantin, and J. Lerga, "Local-entropy based approach for X-ray image segmentation and fracture detection," *Entropy*, vol. 21, no. 4, 2019, doi: 10.3390/e21040338.
- [26] C.-C. Zhang and J.-D. Fang, "Edge Detection Based on Improved Sobel Operator," in *International Conference on Computer Engineering and Information Systems*, 2016, vol. 52, doi: 10.2991/ceis-16.2016.25

A Hybrid Material Generation Algorithm with Probabilistic Neural Networks for Solving Classification Problems

Mohammad Wedyan¹, Omar Alshaweesh², Enas Ramadan³
Ryan Alturki⁴, Foziah Gazzawe⁵, Mohammed J. Alghamdi⁶

Department of Autonomous Systems, Al-Balqa Applied University, Al-Salt, Jordan¹
Department of Software Engineering, AL-Hussein Bin Talal University, Ma'an, Jordan²
Department of Computer Science, Al-Balqa Applied University, Al-Salt, Jordan³
Department of Information Science, College of Computer and Information Systems^{4, 5, 6}
Umm Al-Qura University, Makkah, Saudi Arabia^{4, 5, 6}

Abstract—Classification is based on machine learning, in which each element in a set of data is classified into one of a predetermined set of groups. In data mining, an artificial neural network (ANN) is the most significant methodology because of the exact results obtained through this algorithm and applied in solving many classification problems. ANN consists of a group of types of feed-forward networks, feed-back network, RBF networks, and the probabilistic neural networks (PNN). For classification issues, the PNN is frequently utilized. The primary goals of this research are to fine-tune the weights of neural networks to enhance the classification accuracy. To accomplish this goal, the Material Generation Algorithm (MGA) was investigated with PNN in a hybrid model. Newly, the hybridization of algorithms is ubiquitous and it has led to the development of unique procedures that outperform those that use a single algorithm. Several distinct classification tasks are used to test the efficiency of the suggested (MGA-PNN) approach. The MGA algorithm's efficiency is evaluated using the PNN training outcomes generated, and its outcomes are compared to that of other optimization strategies. By 11 benchmark datasets, the suggested algorithm's performance in terms of classification accuracy is evaluated. The outcomes display that the MGA outperforms the biogeography based optimization, firefly method in terms of classification accuracy.

Keywords—Artificial neural network (ANN); material generation algorithm (MGA); classification; probabilistic neural networks (PNN)

I. INTRODUCTION

Recently, our ability to collect data has greatly improved [1]. Millions of databases have been utilized in a variety of applications, including marketing campaigns, company management, scientific endeavors, and several others [2]. The availability of sophisticated and affordable database systems has resulted in a growing growth in the number of such databases [2, 3]. There is a great need to resort to intelligent approaches to get knowledge from processed data. As a result, data mining has become a popular research field [4, 5]. Classification is a supervised machine learning problem in which a collection of training data is used to map input data into one of several predetermined categories [5, 6]. Any classification algorithm's purpose is to create a technique

which can reliably predict the category of unobserved examples [7]. Classification has numerous uses in a range of areas, including document organization, medical diagnosis, and many more. Many classification techniques and models have been devised and used as a result of this, including radial basis function (RBF) network [8], naive Bayes (NB) classifier [9], support vector machines (SVMs) [10], and K-nearest neighbours (KNN) algorithm [11] and several others.

To solve classification difficulties, ANNs have been frequently used [12]. There are several different kinds of ANNs [13] as modular neural networks, RBF networks, feed-forward neural networks, learning vector quantization neural networks, and several others. Not only do the aforementioned ANNs differ in terms of how they apply to learning, but also in terms of their control method and topology [14].

The PNN is considered a feed-forward neural network that can be used to predict challenges and solve classification. The gradient steepest descent approach, a common optimization technique, is used in the PNN technique to minimize errors between the predicted and actual output functions by allowing the network to modify the weights of the network [15].

The goal of merging metaheuristic algorithms and NN to build classification tools like the PNN is to improve effectiveness and efficiency while also allowing for more accurate and faster solutions of complex problems.

The problem statement was specified by one basic research question: Can the searchability of the Material Generation Algorithm have the ability to choose the best weights so we can get the best accuracy?

Metaheuristics are split into two kinds: population-based and single-based. Genetic algorithm (GA) [16], particle swarm optimization (PSO) [17], water evaporation optimization (WEO) [18], differential evolution (DE) [19], firefly algorithm (FA) [20], artificial bee colony (ABC) [21] and several others are population-based metaheuristics. Local search (LS) [22], tabu search [23] and simulated annealing (SA) [24] are examples of single-based metaheuristics.

The Material Generation Algorithm (MGA) was investigated and utilized in this research to enhance the efficiency of the PNN in solving the classification problem [25]. The PNN was utilized to generate some preliminary solutions that were generated at random and the MGA was utilized to tune the weights of the PNN.

The study is divided as follows. Section II is showed a background and literature review for the MGA. In Section III, the background on big data and its issues is presented. The PNN approach is described in detail in Section IV, while the MGA is described in detail in Section V. The proposed methodology is then detailed in Section VI. In Section VII, the outcomes are presented. Finally, in Section VIII, the conclusion is offered.

II. BACKGROUND AND LITERATURE: MATERIAL GENERATION ALGORITHM (MGA)

The author in [25] suggested that MGA be used to solve engineering challenges in the best possible design. The MGA has identified some of the advanced and fundamental parts of materials chemistry as inspirational concepts, notably the formation of chemical molecules and chemical reactions in the production of new materials. This research demonstrates that the MGA is able to produce highly competitive, if not exceptional, outcomes that outperform other metaheuristics.

The author in [26] presented the optimal design of truss structures using the MGA. For statistical purposes, many optimization runs are carried out. The results showed that the MGA may produce extremely acceptable, resulting in the smallest potential weight compared with the outcomes of a number of metaheuristic methods.

The author in [27] optimized the moulding parameters of resin reinforced sand mould cores using a hybrid Taguchi-WASPAS- MGA to get the optimum outcomes.

The author in [28] used sunflower optimization algorithm and MGA for efficient generation and analysis of materials and equipment of mechanical reducer for the material handling industry. The results showed that the technique is precise in providing better output.

III. BIG DATA: OPPORTUNITIES AND CHALLENGES

The existence of trillions of records that have been produced by millions of people and kept in a variety of online sites suggests the concept of big data [28]. Scientists can use the big data to address issues with small data samples by giving adequate test data to evaluate models, better handling noisy train data, avoiding overfitting models to train data and loosening theoretical model assumptions. In Big Data, there are challenges, like capturing, transferring, storing, cleaning, analyzing, filtering, searching, sharing, securing, and visualizing data [29]. Different research communities have been battling to produce a dynamic, fast, new, and user-friendly Big Data technology [28], which contribute to solving many problems related to data and how to retrieve it.

IV. PROBABILISTIC NEURAL NETWORK (PNN)

The PNN was proposed for the first time in[30]. The training of a PNN does not entail using heuristic searches to

find the best smoothing factor, as this is an optimization problem [31]. A four-layered feed-forward network is formed: (a) input layer, (b) hidden layer, (c) summation layer, (d) output layer, using a statistical algorithm. Fig. 1 illustrates the architecture of a typical PNN. Each input neuron acts as a unique characteristic from the train and test datasets [32]. The PNN network's four levels are detailed below:

- Input layer: Each indicator variable is represented by a neuron. The categorical factors are made up of N-1 neurons, with N being the number of categories. By subtracting the middle value, the input neuron is expected to normalize the value range. It then divides it into quartile range values.
- Hidden layer: Each occurrence in the training dataset is represented by a single neuron. Each training sample has one unit which creates a product of the input vector x and the weight vector wi, zi = x.wi, and then runs the nonlinear procedure:

$$\exp \left[\left(\frac{(w_i - x)^T \cdot (-w_i - x)}{(2\alpha^2)} \right) \right] \quad (1)$$

- Pattern/summation layer: A single pattern neuron is available for each class of objective criteria. The weight value that emerges from the hidden neurons is given to the pattern neurons that match with the hidden neurons. Each training group's objective class is stored alongside each hidden neuron. which combines the contributions for each type of input and provides the output of a network as a probabilistic vector:

$$\Sigma_i \left[\left(\frac{(-w_i - x) \cdot (w_i - x)^T}{(2\alpha^2)} \right) \right] \quad (2)$$

- Output / Decision layer: creates binary classes that correspond to the decision classes Ω_s and Ω_r , $s \neq r$, $s, r = 1, 2, \dots, q$ relies on the following criteria of classification.

$$\Sigma_i \left[\left(\frac{(w_i - x)^T \cdot (-w_i - x)}{(2\alpha^2)} \right) \right] > \Sigma_j \left[\left(\frac{(w_j - x)^T \cdot (-w_j - x)}{(2\alpha^2)} \right) \right] \quad (3)$$

There is just one weight for these nodes, C, the number of training samples in each class and the prior membership probabilities, C, given by the cost parameter:

$$C = \frac{h_s l_s}{h_r l_r} \cdot \frac{n_r}{n_s} \quad (4)$$

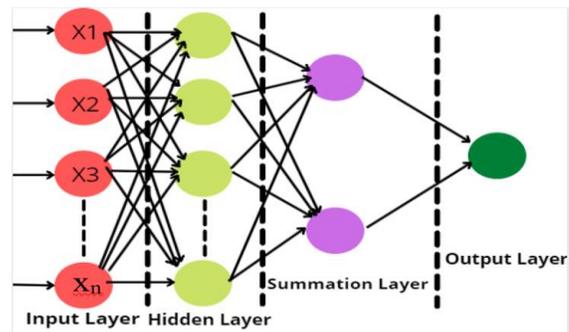


Fig. 1. Probabilistic Neural Network Structure.

V. MATERIAL GENERATION ALGORITHM (MGA)

In the year 2021, MGA is a bioinspired algorithm inspired by material chemistry[25]. To construct and formulate a well-defined mathematical model for the new method, the basic principles of chemical compounds, reactions, and stability are used. MGA determines a number of materials (Mat) made of several periodic table elements (PTEs), based on the fact that such natural evolution technique create a preset population of solution candidates that are evolved by random changes and selection. A materials numbers are examined as solution candidates (Mat_n) in this algorithm, each of which is made up of some constituents that are represented as decision variables (PTE_jⁱ). The following is the mathematical representation of these two components:

$$\text{Mat} = \begin{bmatrix} \text{Mat}_1 \\ \text{Mat}_2 \\ \vdots \\ \text{Mat}_i \\ \vdots \\ \text{Mat}_n \end{bmatrix} = \begin{bmatrix} \text{PTE}_1^1 & \text{PTE}_1^2 & \dots & \text{PTE}_1^j & \dots & \text{PTE}_1^d \\ \text{PTE}_2^1 & \text{PTE}_2^2 & \dots & \text{PTE}_2^j & \dots & \text{PTE}_2^d \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ \text{PTE}_i^1 & \text{PTE}_i^2 & \dots & \text{PTE}_i^j & \dots & \text{PTE}_i^d \\ \vdots & \vdots & \dots & \vdots & \ddots & \vdots \\ \text{PTE}_n^1 & \text{PTE}_n^2 & \dots & \text{PTE}_n^j & \dots & \text{PTE}_n^d \end{bmatrix}, \quad \begin{cases} i=1,2,\dots,n. \\ j=1,2,\dots,d. \end{cases} \quad (5)$$

There are two variables in the mathematical equations where d denotes the number of items (decision variables) in each subject (the candidate solutions) and n denotes the total number of items considered.

PTE_jⁱ is determined at random in the first step of the optimization procedure, whereas the decision variable boundaries are defined based on the problem under consideration. The initial placements of PTEs in the search space are set at random:

$$\text{PTE}_{i,j}^j(0) = \text{Unif}(0,1) \cdot (\text{PTE}_{i,max}^j - \text{PTE}_{i,min}^j) + \text{PTE}_{i,min}^j, \quad \begin{cases} i=1,2,\dots,n. \\ j=1,2,\dots,d. \end{cases} \quad (6)$$

Where PTE_jⁱ(0) is the beginning value of the jth element in the ith material; PTE_{i,min}^j and PTE_{i,max}^j are the minimum and maximum permissible values for the jth decision variable of the ith solution candidate, respectively; and Uni f(0, 1) is a random number in the [0, 1] range.

To mathematically simulate chemical compounds, all PTEs are considered to be in the ground state, which can be externally activated by magnetic areas, photon or light absorption, and interactions with other colliding entities or particles in the case of ions or other individual electrons. Elements have a tendency to gain, lose, or even share electrons with other PTEs due to their varied stabilities, resulting in ionic or covalent compounds. Using the initial Mat in equation (5), d random PTEs are chosen to model the ionic and covalent compounds. The probability theory is used to model the operations of sharing electrons, gaining, and losing for the selected PTEs. To achieve this goal, for each PTE, a continuous probability distribution is used to configure a

chemical molecule, which is then regarded a new PTE, as follows:

$$\text{PTE}_{new}^k = \text{PTE}_{r_1}^{r_2} \pm e^-, \quad k=1,2,\dots,d \quad (7)$$

R2 and r1 are random integers uniformly distributed in the intervals [1, d] and [1, n], respectively; PTE_{r₁}^{r₂} is from the Mat that was chosen at random; e- is the probabilistic component for simulating electron loss, gain, and sharing in the mathematical model represented with a normal Gaussian distribution; and PTE_{new}^k new is the new material. PTEs are used to construct a new material (Mat_{new1}), which is being added as a new solution filter to the list of the raw material (Mat):

$$\text{Mat}_{new1} = [\text{PTE}_{new}^1 \text{ PTE}_{new}^2 \dots \text{PTE}_{new}^k \dots \text{PTE}_{new}^d], \quad k=1,2,\dots,d \quad (8)$$

The candidates for the overall solution are then integrated and displayed as follows:

$$\text{Mat} = \begin{bmatrix} \text{Mat}_1 \\ \text{Mat}_2 \\ \vdots \\ \text{Mat}_i \\ \vdots \\ \text{Mat}_n \\ \text{Mat}_{new1} \end{bmatrix} = \begin{bmatrix} \text{PTE}_1^1 & \text{PTE}_1^2 & \dots & \text{PTE}_1^j & \dots & \text{PTE}_1^d \\ \text{PTE}_2^1 & \text{PTE}_2^2 & \dots & \text{PTE}_2^j & \dots & \text{PTE}_2^d \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ \text{PTE}_i^1 & \text{PTE}_i^2 & \dots & \text{PTE}_i^j & \dots & \text{PTE}_i^d \\ \vdots & \vdots & \dots & \vdots & \ddots & \vdots \\ \text{PTE}_n^1 & \text{PTE}_n^2 & \dots & \text{PTE}_n^j & \dots & \text{PTE}_n^d \\ \text{PTE}_{new}^1 & \text{PTE}_{new}^2 & \dots & \text{PTE}_{new}^k & \dots & \text{PTE}_{new}^d \end{bmatrix}, \quad \begin{cases} i=1,2,\dots,n. \\ j=1,2,\dots,d. \\ k=1,2,\dots,d. \end{cases} \quad (9)$$

Fig. 2 depicts the structure of the mentioned method for configuring new materials based on chemical components (covalent and ionic).

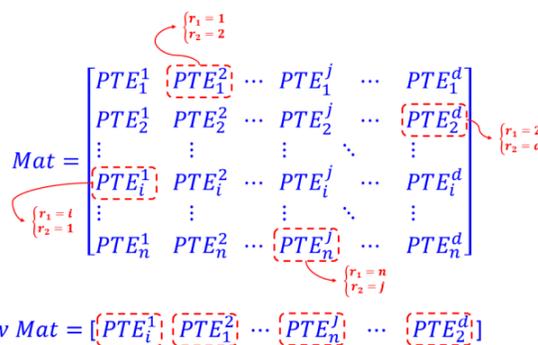


Fig. 2. The Random Periodic Table Elements (PTE) Selection and Creation of New Materials are Depicted Schematically.

The following is the probability of selecting a new element (PTE_{new}^k) in relation to the randomly picked first element (PTE_{r₁}^{r₂}):

$$f(\text{PTE}_{new}^k | \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad k=1,2,\dots,d \quad (10)$$

The symbol for the standard deviation in the previous equation is σ; the symbol for the variance is σ²; μ is the

median, expectation of the distribution or mean, which corresponds to the randomly chosen PTE (PTE_{r1}^2); and e is the natural logarithm's Napierian base or natural base.

Chemical reactions are a type of manufacturing method in which various chemical changes are decided for producing products with altered characteristics that are even distinct from the initial reaction mixture. To simulate the procedure of manufacturing new materials mathematically using the reaction mixture idea, an integer random number (l) is determined depend on the materials in the first Mat that are examined for participation in a reaction mixture. After that, to decide the placements of the picked materials in the initial Mat , l integer random numbers (m_j) are created. As a result, new solutions are created that are linear combinations of the previous ones. Fig. 3 depicts a schematic illustration of the given procedure, with the following mathematical representation:

$$Mat_{new2} = \frac{\sum_{m=1}^l (p_m \cdot Mat_{m_j})}{\sum_{m=1}^l (p_m)}, j=1,2,\dots,l \quad (11)$$

The Mat_m is the m th randomly chosen material from the first Mat , Mat_{new2} is the new material created by the chemical reaction idea and p_m is the normal Gaussian distribution for the m th material participation factor.

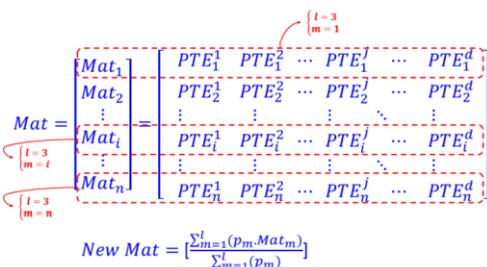


Fig. 3. A Diagram of the Random Material Selection Method for Developing New Materials.

VI. METHODOLOGY PROPOSED: MGA WITH PNN

The MGA was utilized in this research to determine the best weights to employ with the PNN algorithm. To address the classification issues, it suggested the MGA–PNN, a new hybrid method. As shown in Fig. 4, the method begins with the PNN producing the initial weights randomly. Following that, the input values are multiplied by the matching weights w_{ij} , which are based on the PNN model's values.

The proposed MGA–PNN structure is shown in Fig. 5. It is divided into two sections: the first is the PNN, which makes utilize the training data. The data that has been tested is then classified. The accuracy is calculated using equation (12). The MGA is then used to fine-tune the PNN weights. The new data will then be tested for accuracy. This method is continued until the end criteria have been fulfilled.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (12)$$

The object is categorized as TN if both the expected and actual labels are negative. The class is categorized as TP if both the expected and actual labels of the object are positive. Further, the class is categorized as FP, when the anticipated class is positive, but the actual label is negative. The

anticipated class is negative, but the actual label is positive, therefore it's categorized as FN. See Table I [33].

For evaluating the proposed MGA-PNN performance, three additional performance measurements were calculated: The rate of error was found (equation (13)), specificity (equation (14)), sensitivity (equation (15)) and G-mean (equation (16)).

$$Error\ Rate = 1 - \frac{TP+TN}{TP+TN+FP+FN} \quad (13)$$

$$Specificity = \frac{TN}{FP+TN} \quad (14)$$

$$Sensitivity = \frac{TP}{FN+TP} \quad (15)$$

$$G\text{-mean} = \sqrt{(Sensitivity \times Specificity)} \quad (16)$$

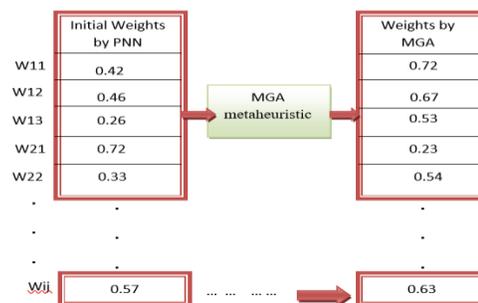


Fig. 4. Representation of Initial Weights.

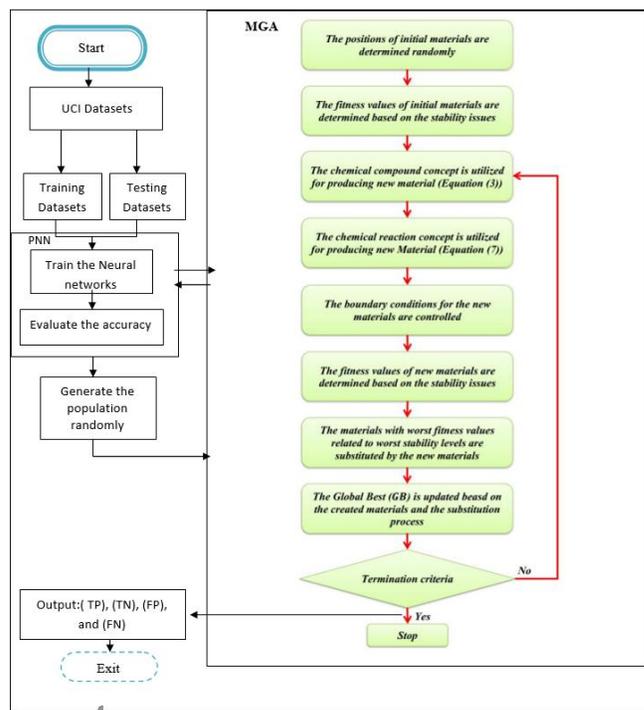


Fig. 5. Flowchart of MGA-PNN Technique.

TABLE I. CROSS-MATRIX CLASSIFICATION

	Positive	Negative
Positive	TP	FN
Negative	FP	TN

VII. EXPERIMENTS AND RESULTS

The efficiency of the MGA-PNN approach is measured in this research using 11 benchmark UCI datasets. We will compare the outcomes of proposed method (MGA-PNN) with PNN, biogeography-based optimization (BBO) and firefly algorithm (FA).

A. Description of the Dataset

These studies are based on a set of a datasets, which may be found at [7]. The previous link provides the size of the testing and training sets. The split was made using a basic train/test split algorithm with a training size is equal to 0.7 and a testing size is equal to 0.3.

B. The Categorization Quality Evaluation Results

Experiments are run on a Windows 10 professional PC with MATLAB R2015b and 16 GB RAM with an Intel ®

Xeon ®CPU ES-1630 v3 @3.70 GHz computer. Table II displays the settings for the input parameters.

The recommended method's rating quality is determined by their ability to improve the desired solution. Table III compares the performance of the proposed MGA technique with PNN, FA [34] and BBO in terms of ratio G-mean, error rate (%), specificity, accuracy and sensitivity.

TABLE II. INPUT PARAMETER SETTING

NCompan (Maximum number of initial Component)	100
Globalbest	0
Max_iteration	200
Population_size	50

TABLE III. CLASSIFICATION SPECIFICITY, ACCURACY, ERROR RATE, RATIO G-MEAN AND SENSITIVITY FOR PNN, FA, BBO AND MGA

Dataset	Technique	TP	FP	TN	FN	Accuracy	Sensitivity	Specificity	Error rate	Ratio g-mean
PID	PNN	35	28	90	39	65.104	47.30	76.27	34.89	60.06
	FA-PNN	33	30	113	16	76.040	67.35	79.02	24.00	72.95
	BBO-PNN	38	25	99	30	71.350	55.88	79.84	28.56	66.79
	MGA-PNN	164	14	43	28	82.80	85.05	75.0	17.2	78.87
HSS	PNN	44	12	6	15	64.93	74.58	33.33	35.07	49.86
	FA-PNN	54	2	10	11	83.12	83.08	83.33	16.88	83.20
	BBO-PNN	52	4	11	10	81.82	83.87	73.33	18.18	78.42
	MGA-PNN	71	3	94	5	95.38	93.42	96.91	4.62	95.15
AP	PNN	23	1	1	2	88.88	92.00	50.00	11.12	67.82
	FA-PNN	24	0	1	2	92.59	92.31	100.00	7.41	96.08
	BBO-PNN	52	4	11	10	81.82	83.87	73.33	18.18	78.42
	MGA-PNN	93	0	7	3	93.88	92.86	100.00	6.12	96.36
BC	PNN	14	9	36	13	69.44	51.9	80.00	30.60	64.44
	FA-PNN	31	1	24	12	80.88	72.09	96.00	19.12	83.19
	BBO-PNN	13	10	44	5	79.17	72.22	81.48	20.83	76.71
	MGA-PNN	17	6	44	5	84.72	77.27	88.00	12.28	82.46
LD	PNN	18	15	34	19	60.46	48.60	69.40	39.50	58.08
	FA-PNN	31	1	24	12	79.07	72.09	96.00	20.93	83.19
	BBO-PNN	32	0	23	13	72.09	71.11	100.00	27.91	84.30
	MGA-PNN	24	9	49	4	84.88	85.71	84.48	15.12	85.09
Heart	PNN	27	5	23	13	73.53	67.50	82.10	26.50	74.44
	FA-PNN	31	1	24	12	80.88	72.09	96.00	19.12	83.19
	BBO-PNN	32	0	23	13	80.90	71.11	100.00	19.10	84.33
	MGA-PNN	32	0	25	11	83.82	74.42	100.00	16.18	86.27
GCD	PNN	133	46	39	32	68.80	80.60	45.90	31.20	60.82
	FA-PNN	166	13	30	41	78.40	80.19	69.77	21.60	74.79
	BBO-PNN	139	40	44	27	73.20	83.73	52.38	26.80	66.23
	MGA-PNN	165	14	42	29	82.80	85.05	75.00	17.20	79.87
Parkinson's	PNN	39	0	4	6	87.75	86.67	100.00	12.25	93.09
	FA-PNN	38	1	6	4	89.80	90.48	85.71	10.2	88.06
	BBO-PNN	39	0	7	3	93.88	92.86	100.00	6.12	96.36
	MGA-PNN	93	0	7	3	93.88	92.86	100.00	6.12	96.36
SPECTF	PNN	49	4	5	9	80.59	84.48	55.56	19.40	68.51
	FA-PNN	52	1	10	4	92.54	92.86	90.91	7.46	91.88
	BBO-PNN	49	9	5	5	86.57	90.74	35.71	13.43	56.92
	MGA-PNN	49	4	12	2	91.04	96.08	75.00	8.96	84.88
ACA	PNN	60	14	84	15	83.24	80.00	85.70	16.8	82.80
	FA-PNN	65	9	94	5	91.91	92.86	91.26	8.09	92.06
	BBO-PNN	65	9	88	11	88.53	85.53	90.72	11.47	88.09
	MGA-PNN	24	9	49	4	84.88	85.71	84.48	15.12	85.09
Fourclass	PNN	59	19	127	11	86.11	84.29	86.99	13.89	85.63
	FA-PNN	78	0	138	0	100.00	100.00	100.00	0.00	100.00
	BBO-PNN	78	0	138	0	100.00	100.00	100.00	0.00	100.00
	MGA-PNN	78	0	138	0	100.00	100.00	100.00	0.00	100.00

The outcomes showed that the proposed algorithm is superior in 9 out of 11 datasets over the rest of the algorithms in Table III. The original PNN achieved 65.1% accuracy in the PIMA Indian diabetes (PID) dataset, while the proposed MGA-PNN attained 82.8 percent accuracy. All the best outcomes showed in bold. The suggested technique has strong exploitation capabilities and can come up with superior solutions because a large number of candidates are grouped around the best solution. On almost all datasets, the suggested MGA outperforms the original PNN approach in terms of error rate, sensitivity, specificity and accuracy.

The MGA's performance was further validated by examining whether it differed statistically from the FA. For classification accuracy, a t-test with a significance interval of 95 percent ($\alpha = 0.05$) was used. Table IV displays the suggested approach's standard deviations and accuracy means. The performance of the MGA is clearly superior to that of the FA, as all of the P-values are less than 0.01.

TABLE IV. THE P-VALUES FOR MGA ACCURACY WITH FA AND T-TEST ACCURACY

Dataset		Mean	Std.Deviation	Std.Error Mean	P-Value
PID	MGA	82.8000	0.00000	0.00000	0.00
	FA	73.4895	1.28560	0.23472	
HSS	MGA	94.5087	0.75519	0.13788	0.00
	FA	81.8179	1.02322	0.18681	
AP	MGA	92.5170	1.72282	0.31450	0.00
	FA	92.5926	0.00012	0.00002	
BC	MGA	82.9167	1.59613	0.29141	0.00
	FA	77.3935	1.74347	0.31831	
LD	MGA	83.3333	2.23006	0.40720	0.00
	FA	75.5810	1.49604	0.27310	
Heart	MGA	82.2549	1.80867	0.33022	0.00
	FA	78.6819	2.23781	0.40857	
GCD	MGA	82.8000	0.00000	0.00000	0.00
	FA	75.1600	1.58040	0.28854	
Parkinson's	MGA	92.5170	1.72282	0.31450	0.00
	FA	89.7950	0.00000	0.00000	
SPECTF	MGA	88.2668	1.37125	0.25035	0.00
	FA	88.8057	1.82787	0.33372	
ACA	MGA	83.3333	2.23006	0.40720	0.00
	FA	89.8840	1.05983	0.19350	
Fourclass	MGA	100.000	0.00000	0.00000	0.00
	FA	100.000	0.00000	0.00000	

VIII. CONCLUSION

The major goal of this study was to propose a new strategy for determining high-quality answers to categorization problems. The Material Generation Algorithm is a population-based metaheuristic that MGA is a bioinspired algorithm inspired by material chemistry. Therefore, the weight values of

the PNN can be optimized by MGA. When a huge search space is being examined, the MGA's superior exploitation and exploration capabilities allow it to achieve better results than FA and BBO. The MGA was utilized to tune the weight of the PNN in this study. To attain the research's targets, the results of this strategy rely on PNN and MGA was used to compare with the results original PNN's classification accuracy, FA-PNN and BBO-PNN. The MGA, which optimized the PNN weights, was used to improve the initial solutions, which were created randomly using the PNN. According to experimental results utilizing 11 benchmark datasets, the suggested MGA with PNN outperformed the original PNN, FA-PNN and BBO-PNN on 9 out of 11 benchmark datasets. This leads us to the fact that MGA can be implemented in additional real and high dimensional datasets to investigate their behavior under different situations in terms of trait numbers. As a result, we'll be focusing our efforts on this topic in the future.

IX. DISCUSSION

This study is considered one of the most important studies in the world of Data Mining. As our use of the method of merging with the metaphysical algorithms, especially with the MGA, and comparing its results with the results of 3 other studies (PNN, FA and BBO) that gives clear evidence of its importance in terms of increasing classification accuracy. As this study only used one algorithm to combine it with PNN, I believe that merging more than one of the high-specification meta-historical algorithms with PNN leads to an improvement and a significant increase in accuracy, and this is our destination in the work of these studies in the future.

REFERENCES

- [1] M. O. Wedyan, "Augmented reality and novel virtual sample generation algorithm based autism diagnosis system," 2020.
- [2] K. Khatatneh, I. Khataleen, R. Alshwaiyat, and M. Wedyan, "A Novel Student Clustering Model for the Learning Simplification in Educational Environments," vol. 10, no. 8, 2019.
- [3] J. Atwan et al., "The effect of using light stemming for Arabic text classification," vol. 12, no. 5, 2021.
- [4] M.-S. Chen, J. Han, and P. S. Yu, "Data mining: an overview from a database perspective," vol. 8, no. 6, pp. 866-883, 1996.
- [5] A. Khtoom and M. Wedyan, "Feature Selection Models for Data Classification: Wrapper Model vs Filter Model," in International Conference on Information, Communication and Computing Technology, 2019, pp. 247-257: Springer.
- [6] M. Wedyan, A. Al-Jumaily, and A. Crippa, "Using machine learning to perform early diagnosis of autism spectrum disorder based on simple upper limb movements," vol. 15, no. 4, pp. 195-206, 2019.
- [7] M. Wedyan, H. Al-Zoubi, and J. Atwan, "Error Detection and Correction Using a Genetic Algorithm," vol. 8, no. 2, pp. 9-16, 2019.
- [8] R. J. Howlett and L. C. Jain, Radial basis function networks 2: new advances in design. Physica, 2013.
- [9] S. Xu, "Bayesian Naïve Bayes classifiers to text classification," vol. 44, no. 1, pp. 48-59, 2018.
- [10] I. Steinwart and A. Christmann, Support vector machines. Springer Science & Business Media, 2008.
- [11] O. Sutton, "Introduction to k nearest neighbour classification and condensed nearest neighbour data reduction," vol. 1, 2012.
- [12] M. Wedyan, A. Al-Jumaily, and A. Crippa, "Early diagnose of autism spectrum disorder using machine learning based on simple upper limb movements," in International conference on hybrid intelligent systems, 2018, pp. 491-500: Springer.

- [13] M. Wedyan, A. Crippa, and A. Al-Jumaily, "A novel virtual sample generation method to overcome the small sample size problem in computer aided medical diagnosing," vol. 12, no. 8, p. 160, 2019.
- [14] C. Gershenson, "Artificial neural networks for beginners," 2003.
- [15] G. P. Zhang and P. C. Cybernetics, "Neural networks for classification: a survey," vol. 30, no. 4, pp. 451-462, 2000.
- [16] G. R. Harik, F. G. Lobo, and D. E. Goldberg, "The compact genetic algorithm," vol. 3, no. 4, pp. 287-297, 1999.
- [17] F. Marini, B. J. C. Walczak, and I. L. Systems, "Particle swarm optimization (PSO). A tutorial," vol. 149, pp. 153-165, 2015.
- [18] A. Kaveh, T. Bakhshpoori, and Structures, "Water evaporation optimization: a novel physically inspired optimization algorithm," vol. 167, pp. 69-85, 2016.
- [19] K. V. Price, "Differential evolution," in Handbook of optimization: Springer, 2013, pp. 187-214.
- [20] X.-S. Yang, "Firefly algorithms for multimodal optimization," in International symposium on stochastic algorithms, 2009, pp. 169-178: Springer.
- [21] F. S. Abu-Mouti and M. E. El-Hawary, "Overview of Artificial Bee Colony (ABC) algorithm and its applications," in 2012 IEEE International Systems Conference SysCon 2012, 2012, pp. 1-6: IEEE.
- [22] V. Kalogeraki, D. Gunopulos, and D. Zeinalipour-Yazti, "A local search mechanism for peer-to-peer networks," in Proceedings of the eleventh international conference on Information and knowledge management, 2002, pp. 300-307.
- [23] F. Glover and M. Laguna, "Tabu search," in Handbook of combinatorial optimization: Springer, 1998, pp. 2093-2229.
- [24] P. J. Van Laarhoven and E. H. Aarts, "Simulated annealing," in Simulated annealing: Theory and applications: Springer, 1987, pp. 7-15.
- [25] S. Talatahari, M. Azizi, and A. H. Gandomi, "Material generation algorithm: a novel metaheuristic algorithm for optimization of engineering problems," vol. 9, no. 5, p. 859, 2021.
- [26] M. Azizi, M. B. Shishehgarkhaneh, and M. Basiri, "Optimum design of truss structures by Material Generation Algorithm with discrete variables," vol. 3, p. 100043, 2022.
- [27] N. C. Behera, S. Jeet, C. K. Nayak, D. K. Bagal, S. N. Panda, and A. Barua, "Parametric appraisal of strength & hardness of resin compacted sand castings using hybrid Taguchi-WASPAS-Material Generation Algorithm," vol. 50, pp. 1226-1233, 2022.
- [28] S. Jena, S. Jeet, D. K. Bagal, A. K. Baliarsingh, D. R. Nayak, and A. Barua, "Efficiency analysis of mechanical reducer equipment of material handling industry using Sunflower Optimization Algorithm and Material Generation Algorithm," vol. 50, pp. 1113-1122, 2022.
- [29] R. Bedeley and L. Iyer, "Big Data opportunities and challenges: the case of banking industry," in Proceedings of the Southern Association for Information Systems Conference, 2014, vol. 1, pp. 1-6.
- [30] D. F. Specht, "Probabilistic neural networks," vol. 3, no. 1, pp. 109-118, 1990.
- [31] M. Wedyan, O. Elshaweesh, E. Ramadan, and R. Alturki, "Vibrating Particles System Algorithm for Solving Classification Problems," vol. 43, no. 3, pp. 1189-1206, 2022.
- [32] W. Sweeney, M. Musavi, and J. Guidi, "Probabilistic neural network as chromosome classifier," in Proceedings of 1993 International Conference on Neural Networks (IJCNN-93-Nagoya, Japan), 1993, vol. 1, pp. 935-938: IEEE.
- [33] M. Alweshah, E. Ramadan, M. H. Ryalat, M. Almi'ani, and A. Hammouri, "Water evaporation algorithm with probabilistic neural network for solving classification problems," vol. 6, no. 1, pp. 1-15, 2020.
- [34] M. Alweshah and S. Abdullah, "Hybridizing firefly algorithms with a probabilistic neural network for solving classification problems," vol. 35, pp. 513-524, 2015.

IoT based Portable Weather Station for Irrigation Management using Real-Time Parameters

Geeta Ambildhuke, Barnali Gupta Banik

Department of Computer Science & Engineering

Koneru Lakshmaiah Education Foundation

Deemed to be University, Hyderabad Telangana-500075, India

Abstract—Rainfall in India is very unpredictable and is characterised by monsoon gaps. Rainfall prediction is very crucial for irrigation management to enhance farm productivity.. This article presents a portable rainfall prediction device which can be carried to fields. In the field by sensing the current atmospheric parameters like temperature, humidity, atmospheric pressure along with the current status of the sky to know the types of clouds present and gives the chances of rainfall. It is a novel approach in terms of portability of the device and it will give the prediction based on current information at a particular location by combining the predictions from the model of image processing of the clouds using deep learning and the currently sensed weather parameters are processed using machine learning without using WIFI or internet connection by providing Edge analytics where the data processing, rainfall prediction, and decision making is carried out locally on the device without any backend servers or cloud platform which will be very useful for the people like farmers who don't have accessibility to internet in villages. The farmers can decide before every irrigation schedule, based on the prediction to what extent the crops can be irrigated. If chances of rain are very low 90% irrigation can be carried out, If chances of rain are predicted as low to medium then 40 to 60% irrigation can be done and if the prediction says medium to heavy rainfall then no irrigation is recommended.

Keywords—Deep learning; edge analytics; internet of things; machine learning; irrigation management; precision agriculture; rainfall prediction

I. INTRODUCTION

The agriculture industry is the backbone of the Indian economy, accounting for roughly 15% of national GDP and about half of India's population is entirely or partially reliant on agriculture and related activities for their living. India is one of the top 15 agricultural commodity exporters in the world. India is also facing an exponential population growth that demands a great need for food in the future. Traditional agricultural practices, on the other hand, must be modified with the help of technology to meet the increasing demand for good quality food to meet the future need of the population. Agriculture development will benefit not just farmers, but also a huge portion of the people living in rural areas that are actively involved in agriculture or indirectly tied to agriculture as consumers. More efficient production methods would produce a more conjugative environment in the country for the overall development of the economy and greater agricultural income. Small and marginal farmers will be empowered via education, reforms, and development,

resulting in better, more efficient, and stronger Indian agriculture. New production and marketing models, as well as raising awareness and providing education to farmers in villages, will aid in the sector's development and, more crucially, improve the economic situation of impoverished farmers [1].

Precision agriculture is the most promising approach to farm management in today's era, as it uses technology Internet of Things (IoT) to monitor the spatial and temporal parameters of the field and ensures that input resources such as water, fertilizers, pesticides, and other chemicals are applied to the crops at the right time, in the right place, and the right amount. Traditional agricultural surveillance tactics were modified by IoT-based agriculture applications, which quickly provided quantitative data with great temporal and spatial resolution. Ecological data is collected in real-time from the surroundings of the agriculture field using various sensors installed in the fields, which will be communicated and analyzed to determine various problems. This collected data is analyzed and processed to extract the proper information and is utilized to decide on various tasks performed in the field to automate the agriculture process to overcome certain difficulties. Multidisciplinary methods that combine remote sensing, modelling, and deep learning techniques can aid in the improvement of agricultural operations and, as a result, crop output [2].

Collecting data manually from a big field is challenging, and it results in fluctuation when compared to inaccurate field measurements. Wireless sensor networks (WSN) help in integrating various sensors in the field for the collection of different parameters from the field and environment. Various farm monitoring applications like crop selection according to the soil in the farm, soil monitoring to make zones with similarity to promote multi cropping and to assist farmers with various decision making like when and in what amount of fertilizer to be used, when to irrigate the crop, early disease predictions etc. are possible due to several emerging technologies like WSN, Deep learning, Machine learning, IoT etc.

Agriculture utilizes 85 percent of the world's available freshwater resources, and this fraction is progressively rising in parallel with population expansion and rising food demand. As a result, there is a need to develop more efficient methods to ensure that water resources are properly utilized in irrigation. New innovative plans in this direction are needed in

research to boost productivity and water management through novel irrigation techniques [3].

Proper irrigation aids in the growth of crops, landscape maintenance, and revegetation of disturbed soils. In addition to frost protection, weed suppression in grain fields, and soil consolidation prevention, irrigation plays important role in healthy crop cultivation. Irrigation management is required even in places with abundant rainfall to boost farm productivity. Monsoon gaps are a feature of Indian rainfall. As a result, sometimes it won't rain for two or more weeks during the rainy season, causing agricultural damage in the absence of irrigation.

In the near future, proper integration of advanced agricultural practices based on various technologies and adaptation by rural communities should be secured. Climate change will have extensive consequences in the next decades. As a result, laws, and practices must be devised to protect farming communities, particularly small landholders, from the immediate losses caused by extreme events [4].

The most important reasons for India's need for perfect irrigation are:

- Chaotic nature of climate.
- Uneven Rainfall Distribution.
- Optimized use of water resources to meet crop needs and soil requirements.
- To maximize crop production.
- To manage supplement supply even in areas with abundant rainfall.

Irrigation scheduling is managed by rainfall forecast based on weather conditions, which determines the amount of water to be supplied by irrigation to maintain the threshold value of water (moisture) required by the crop at any given time.

Automatic irrigation scheduling systems have replaced manual irrigation based on soil water monitoring. While implementing, the plant evapotranspiration was taken into account, which was based on many climatic characteristics such as humidity, wind speed, solar radiation, and even crop aspects such as stage of growth, plant density, soil attributes, and pest.

Weather forecasting does an excellent job of providing people with early warnings and alerts about severe climate change events, however, due to a lack of infrastructure, people in remote regions such as villages do not receive information on time due to the absence of internet, and forecasting is not very accurate for any given location. As a result, this portable device may be taken anywhere and used to anticipate the state of rainfall by sensing current atmospheric characteristics and present sky status to identify clouds related to rainfall.

Most of the farmers are living in rural areas and are not able to access the internet and weather forecasting applications at remote locations like fields. This article presents a portable weather station that uses a novel approach that predicts the intensity of rainfall as No Rain to very Low Rain, Low to Medium Rain or, Medium to high Rain based on the combined

approach of identifying types of clouds present in the sky and some important weather parameters such as temperature, humidity, and atmospheric pressure using affordable sensors that are within the reach of farmers and will work independently of WIFI or internet.

This device consists of Raspberry pi with an integrated camera for taking digital images of the sky to know the type of clouds present in the sky along with some atmospheric parameter sensors like DHT11 sensor, BMP 180 etc. to sense temperature, humidity, and atmospheric pressure and will give the prediction on rainfall by both approach from cloud status and sensed atmospheric parameters. Improvement in crop production and planning by using local climate prediction without the internet can definitely help an individual farmer to increase his yield by managing the irrigation cycle by avoiding irrigating the crop with extra water or less water.

A. Novelty

The proposed work is distinguished from previous work by its multi-modal integration of two separate techniques. The majority of previous research has proposed solutions in irrigation management by providing automatic irrigation based on either atmospheric parameters collected from nearby weather stations / using sensors or only by capturing sky images, but the proposed system takes advantage of both approaches where real-time atmospheric parameters on the fields are sensed using sensors and the sky image at that location is captured to provide better rainfall predictions without using WIFI or internet so that it can be used by farmers or people at remote locations. Due to the global shortage of clean water resources, it is critical that they be utilised to their full capacity, which may be accomplished by utilising water resources wisely and paying equal attention to irrigation water management. This device will help farmers to take decision before every irrigation scheduled cycle to what extent the field should be irrigated by predicting the intensity of rainfall as low, medium or high.

The rest of the article is arranged as follows: Section 2 describes relevant work, Section 3 discusses the proposed system in detail, Section 4 discusses methods and methodology, Section 5 discusses findings, and Section 6 discusses conclusion and future work.

II. RELATED WORK

Weather conditions play an important role in both irrigation requirements and crop performance. Researchers used various weather parameters and different technologies and algorithms to predict the weather to know the status of climate to control the use of water in Irrigation.

The evapotranspiration of water in the soil is affected by temperature and humidity. The thermic level of the atmosphere is described as air temperature, which is generally measured in Celsius, Fahrenheit, or Kelvin degrees. It is the most closely observed weather metric. The presence of water vapour in the air is known as humidity. The proportion of water vapour in the air is expressed in percentage (%). The other major parameters considered by the researchers are luminosity and is defined as the brightness or intensity of light. Lux is the unit of measurement. It causes more water

loss from the soil as the temperature rises with the direct solar radiation. The most important parameter is the rainfall or amount of precipitation that influences whether or not irrigation is necessary as well as the amount of water to be used if irrigation is required. The author in [5] presents a detailed survey on the IoT systems and recent sensors used in an automatic smart irrigation system in Precision Agriculture and discussed various weather monitoring parameters, sensors, soil characteristics. The author in [6] proposed an automated irrigation system which starts the motor and water is supplied only if the level of soil moisture goes below the threshold and the amount of water supply is managed by considering the weather parameters temperature and humidity sensed by the sensor to predict the type of climate as sunny (no rain), cloudy (rain chances <50%) or rainy (rain chances>50%). In [7] to add intelligence to the existing idea of automatic irrigation systems, an autonomous irrigation system is proposed that employs machine learning and predictive algorithms to predict the status of rainfall using historical climate data for the amount of water to be used by calculating the time of motor to be ON based on the soil moisture value for each zone separately. Work proposed in [8] is the development of an efficient IoT architecture that monitors soil, microclimate, and water parameters, as well as performing proper irrigation management based on challenges studied in irrigated farmland of Ethiopia, Kenya, and South Africa. For educated decision making and effective operation of the irrigation management system, indigenous agricultural and expert knowledge, local climate information, particular features of crop and soil are provided to the system. Broadband connection and cloud services are either unavailable or too costly in Sub-Saharan Africa. To overcome these constraints, data processing, network administration, irrigation choices, and farmer communication are all done locally, with no back-end servers involved. In the article [9] the technique was evaluated using five crops at four European sites with varying weather circumstances to optimize irrigation water consumption by taking into consideration soil water availability, local weather predictions, crop physiological condition, and water demands in real-time. Main observations are the major effect of inaccurate predictions of the forecast, unless the target yield was excessively high, was that the target yield was not attained, but the irrigation schedule stayed near to optimum, according to the data. In other words, while the actual yield did not match the objective, just a little amount of irrigation

water was lost. In [10] author proposed an idea to automate the irrigation process to manage the motor's pumping by taking into account the soil moisture content and rainfall forecast. Soil moisture data is continuously sensed using a soil moisture sensor and is sent to raspberry pi and depends upon the value a rainfall prediction is obtained from Openweather API to decide on irrigating the field or not and the Android Application is used to track the complete irrigation process for the agricultural field. The article [11] effectively shows a simple, low-cost, and somewhat accurate system for monitoring current meteorological conditions and forecasting rain. This system has a significant benefit over comparable Arduino-based weather monitoring systems in that it also provides rain probability at the current moment using weather parameters namely temperature, humidity, light intensity, and wind conditions. The goal of this study [12] was to provide a way for a reliable irrigation system based on a suitable rainfall forecast algorithm. The new approach, which is Romyan's method is introduced to calculate water requirement for crops, as well as the time necessary for the motor to be turned on. The author in [13] presents a technique for high-accuracy rainfall forecasting that combines two types of data i.e., cloud imaging data and humidity as an atmospheric parameter in numerical form. Convolution Neural Network is used for image recognition by extracting important features using ResNet. In this research, a novel network is built that combines the cloud picture with other meteorological information and creates the result, to forecast rainfall with greater accuracy than the existing ResNet image recognition alone.

A comparison of the existing and proposed work based on the parameters and technology along with methods and methodologies used are shown in Table I. The comparative analysis of the previous research work has revealed that most of the models used either atmospheric parameters or cloud images for the prediction of rainfall and only one paper proposed a system using a combined approach but using only one parameter as humidity. However, after researching existing systems it is clear that the combined approach of two modalities to predict rainfall gives more clarity and would be very useful. In this regard, the system proposed uses a multisource data approach that combines predictions from digital cloud images and atmospheric parameters to provide an improved rainfall prediction system.

TABLE I. COMPARISON OF THE PROPOSED SYSTEM WITH EXISTING SYSTEMS

Model	Atmospheric Parameters used	Source of weather data	Sky status	Methodology used	Output
[6]	temperature and humidity sensors	Sensor deployed in the field	No	Controllers, sensors, and algorithm	Climate predicted as sunny, cloudy, or rainy
[11]	temperature, humidity, light intensity, and wind conditions	Sensors	No	Matlab, Arduino IDE	Automatic Irrigation using soil moisture and rainfall prediction
[8]	Soil and microclimate parameters	sensors	No	IoT, GPRS/GSM communication	autonomous actuations for smart irrigation management
[14]	Temperature, humidity, pressure, and uv_index	Sensors	No	algorithms like SVM, KNN, ANN, etc. explored	Gives rainfall prediction for the next four days
[13]	Humidity	Sensors	yes	CNN, RESNET, and Neural network	rainfall forecasting with cloud imaging data and humidity
<i>Proposed System</i>	<i>Temperature, Humidity, and pressure</i>	<i>Sensors</i>	<i>yes</i>	<i>Transfer learning, IoT, Machine learning</i>	<i>Rainfall intensity as No rain, Low rain, and High rain</i>

The suggested system addresses the optimal integration of different sensing modalities as well as their practical execution. Additionally, machine learning approaches and deep learning architectures are applied to the different inputs and the final decision is provided by combining the output from both approaches to predict rainfall and gives an appropriate decision on the amount of irrigation to be done. In comparison to current procedures, the suggested approach has presented a technology-based solution that would be beneficial to the agricultural and scientific communities in terms of its portability and use of edge analytics where input is processed and output is given at the device level only without cloud platforms, internet or WiFi.

III. PROPOSED SYSTEM

The study of clouds and their characteristics is crucial for a wide range of applications. It's been utilized to produce precise weather forecasts via nowcasting. There exist many clouds in the sky but only a few clouds are responsible for rain. Identification of rainclouds can be done using image classification with the help of deep learning on a ground-based image cloud dataset. To make the predictions more accurate the atmospheric parameters like temperature, humidity, atmospheric pressure is taken into account which can be easily sensed by the sensors and are affordable. The results from both approaches are combined and the final output is given as the prediction on the rainfall as No rain, Low to Medium rain, or heavy rain as shown in Fig. 1 as a multi-modal proposed system on basis of which many agriculture-related activities can be carried out like managing irrigation by taking in account the status and intensity of rainfall.

A. Components used

1) **BMP180:** BMP180 is a barometric sensor with an I2C ("Wire") interface, used to measure pressure surrounding it and altitude. It is also used to measure temperature. It works on push sensor BMP180 and can measure the pressure in the range of 300 to 1100 hPa with relative pressure error to 0.12 hPa (1m height). The BMP180 outputs absolute pressure in pascals (Pa). By observing variations in pressure short-term weather changes can be predicted. Dropping pressure, for example, frequently indicates the arrival of rain or a storm (a low-pressure system is moving in). When the pressure rises, it usually signals that clear weather is on the way (a high-pressure system is moving through).

2) **DHT11** sensor is a basic and commonly used to record temperature and humidity from the atmosphere in the digital form. This sensor utilizes a thermistor and a capacitive humidity sensor to detect the surrounding air. To monitor humidity and temperature instantly, it may be simply interfaced with any microcontroller such as Arduino, Raspberry Pi, and so on.

3) **Raspberry Pi 4:** Raspberry Pi 4 is the latest model and is a tiny processor or a controller with great processing power integrated with Broadcom 2711, 64-bit quad-core Cortex-A72 processor and is available with 1 GB, 2 GB, or 4 GB RAM. It features a true gigabit Ethernet port, 2 x USB 3.0 "Super-Speed" ports which can be used to attach mouse and keyboard. It comes as a size of the credit card so is portable and easy to carry. New version comes with a combination of small footprint, low-power drop, customization and amazing community support and the pi can be used in several.

4) **USB camera:** Logitech HD webcam c270 has a USB interface that makes it easy to connect and has 3-megapixel image resolution along with superb color-rich imaging even in ultra-low light with HD support.

5) **OLED Display Screen:** To display the output as the atmospheric parameters along with the status of rainfall and the decision based on the rainfall is displayed on an OLED display screen. The OLED display screen is very thin and light-weighted with a size of 0.96 inches comes with a resolution of 128X64. It has 4 pins named VCC: 3.3-5V GND: Ground SCL: Serial Clock SDA: Serial Data to carry out I2C communication.

6) **Basic Shield:** Basic Shield is a component provided with 8 LED's, 2 push buttons and is very popular for interfacing electronics components like a push button, potentiometer, LDR, buzzer, etc. and can be easily connected with 5 V/3.3 V microcontrollers.

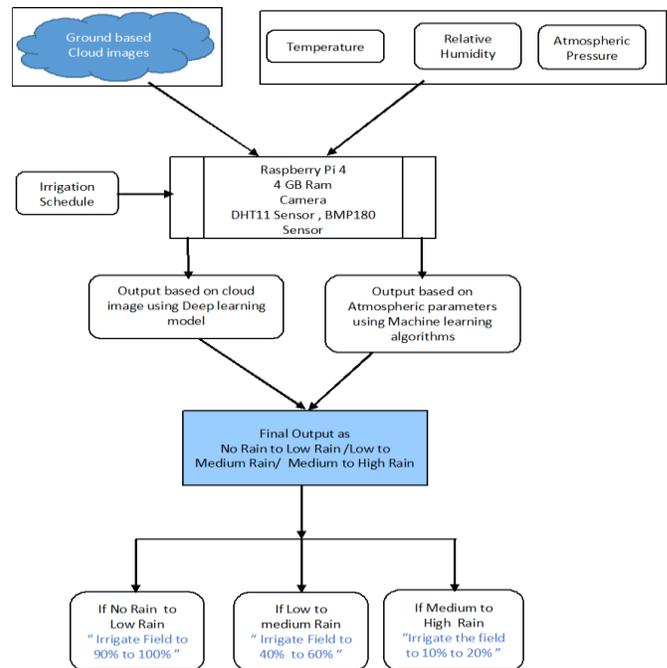


Fig. 1. Proposed System.

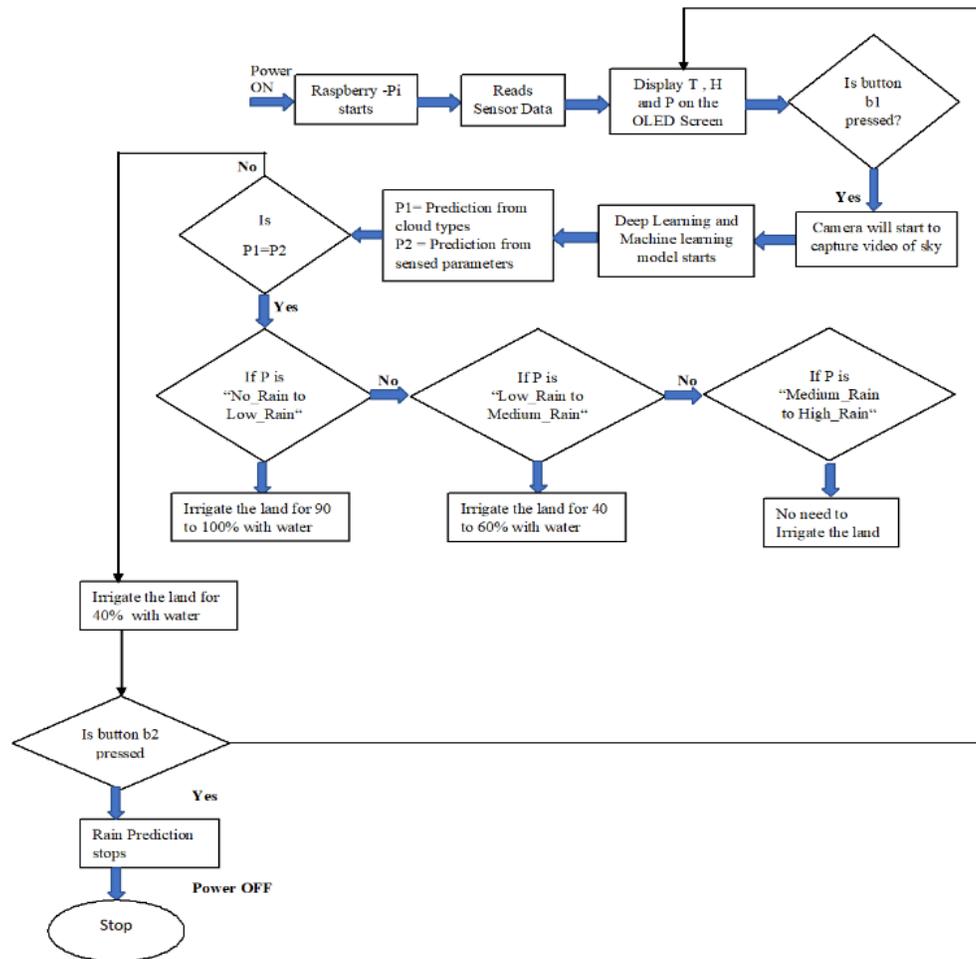


Fig. 2. Working Flow of the Device.

B. Working Principle of the Device

The Device consists of Raspberry Pi 4 Model B and components DHT 11, BMP 180, OLED display screen, USB camera, and Basic-Shield are connected to the respective pins. Power is provided to Raspberry Pi using a power bank with a C-Type cable to make the device portable. The work is shown in Fig. 2. Sensors sense the atmospheric parameters like temperature denoted by T, Humidity denoted by H, and Pressure is denoted by P. Once power is supplied to the device the code snippet for collecting sensor data will start and will get displayed on the screen after regular intervals. To start the Rain prediction push-button b1 needs to press. Once b1 is pressed camera will start and the video will be captured, so the camera should be faced towards the sky to get the proper input, which is fed as input to deep learning model at the same time machine learning code will also be executed by giving values T, H, and P as input parameters.

IV. METHODS AND TECHNOLOGY

A. Irrigation Scheduling

Irrigation scheduling is critical for ensuring that crops receive the right quantity of water at the right time to

minimize crop water stress and optimize output [15]. Water stress can affect vegetable crops in two ways: when there is a lack of water (drought stress) or when there is an abundance of water (water stress) caused due to waterlogging or soil water saturation. A water deficit occurs when water is supplied at the wrong time or insufficiently, reducing the amount of water available to plants in the soil. Long durations of irrigation or high-water application rates can produce excess soil moisture, wasting water and causing nutrient loss. Crop water stress can also have an impact on crop management. Wilting occurs when soil moisture falls below a certain level, preventing plants from drawing water into their roots. Furthermore, in low moisture circumstances, any moisture-activated herbicides and nutrients would not be efficiently used by plants. Presently, many different approaches are adopted for irrigation scheduling as listed in Table II to optimize the water application to crops. These methods are ranked according to the level of management required for water application [16]. Irrigation scheduling is different for a different crop at every growth stage and must be followed to optimize the crop yield and the use of water.

TABLE II. IRRIGATION SCHEDULING METHODS

Rank	Method	Type of water management
0	The "Irrigate whenever" method	Water is applied without scheduling
1	"Feel and appearance" method	The irrigation manager decides the amount of water to be applied and when by observing the soil sample and assessment is done by comparing it with the soil reference images.
2	Systematic method	Regardless of considering weather or soil water conditions, the application of water is done based on the amount or time.
3	Crop water demand method	The amount of water applied is determined by the crop's evapotranspiration (ETc). This strategy involves calendar-based scheduling based on prior seasons and should take into account rainy days.
4	Soil water status method	Water is supplied to the crop root system based on soil moisture levels, usually by giving a proportion of soil accessible water. Rainfall occurrences should be taken into consideration with this technique.
5	Water budgeting method	Water application is dependent on crop evapotranspiration, soil moisture content at the root level, and water budgeting.

The first method is completely traditional and requires human interventions and results in lots of water wastage in terms of irrigation as management was not done properly and no parameters are considered. The second and third methods considered soil parameters and timings as well as an amount for application of water which reduced the water wastage but may result in water stress as no weather conditions are considered. The last three methods are improved and make use of technology by considering various parameters like soil moisture content and rainfall events to plan the irrigation scheduling thus optimizing the use of water resources. From the above table, we can see that Irrigation scheduling is very important to optimize the water application thus by improving the yields and rainfall is one such important factor that must be taken into account to protect the crop from the application of excess water thus improving the Irrigation Management system.

B. Internet of Things (IoT)

Every element of traditional farming processes may be substantially altered by incorporating the newest sensor and IoT technology into agricultural practices. Currently, Wireless sensors and the Internet of Things (IoT) are being integrated into smart agriculture and are capable of taking agriculture sector to next level. Soil, humidity, wetness, light, air temperature, CO2, solar energy sensors, and a variety of other IoT sensors are all employed in agriculture. Sensors, which are placed across the fields, on smart agriculture vehicles, in IoT-based monitoring systems, and weather stations, collect data in real-time and give farmers visibility and control over their activities. Fig. 3 shows the basic architecture of IoT.

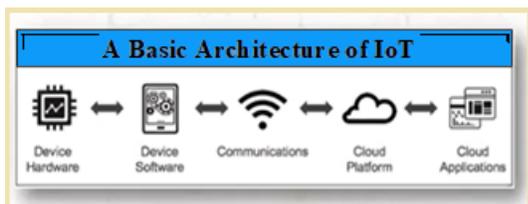


Fig. 3. A Basic Architecture of IoT.

IoT architecture consists of various hardware components like sensors as per the requirement to sense the data from surroundings which requires software and various communication technologies to exchange the data among the devices and finally, the data sensed by the interrelated devices

gets stored on the cloud for further analysis and various application are developed to convert this data into important information required for many useful or decision-making activities.

The information about IoT technology stack that contains companies and produce different level IoT boards or controllers, then come communication technologies NB-IoT, LoRaWAN, ZigBee, Bluetooth, etc. supported by many Communication protocols such as MQTT, CoAP, AMQP, and many more as shown in Fig. 4.

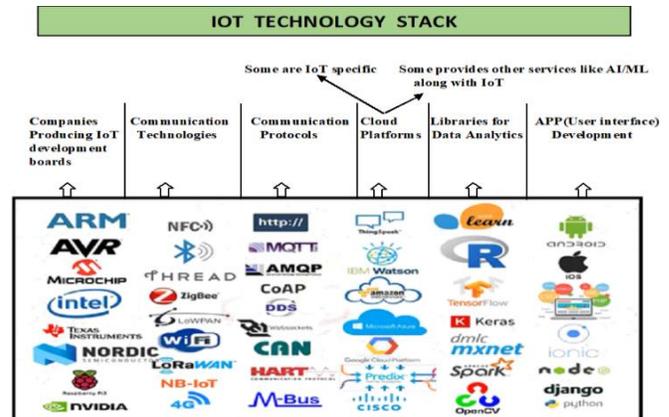


Fig. 4. IoT Components and Technologies.

Today we have clouds like Thingspeak, IBM Watson specifically for IoT along with general-purpose cloud platforms (AWS Amazon, Google, Cisco). Once the data is stored it is analyzed with analytical tools like keras, Tensorflow, OpenCV is libraries with very powerful functions and finally, the information retrieved after analysis of data is used for various applications.

C. Edge Analytics

Sensors are used to collect data from the environment or surroundings and Actuators are the devices that will take action based on the output of processed data. If the Analytics is performed at the device level that concept is called Edge Analytics. Only data for storage is sent to the cloud. If analytics is done in the cloud, then it is called cloud analytics. This experiment used DHT11 and bmp180 sensors to sense temperature, Relative humidity, and atmospheric pressure, and the data is fed to a machine learning algorithm (pickle file) to

get the status of rainfall based on these real-time parameters. Along with it, raspberry pi is integrated with a digital camera to capture the image of the sky at that location, and the image is given as the input to the deep learning model (.h5 file) to get the status of rainfall based on the types of clouds present in the sky at that time. The processing is done in Raspberry pi and the predictions are displayed on the screen attached to it. This proposed device uses Edge analytics to give output as the processing is done at device level and does not depend on WIFI or internet. The cloud can be used to just store the data collected by the sensor only when Wi-Fi is available. The cost of data storage and management is reduced using edge analytics. It also saves operational costs, bandwidth requirements, and time spent on data analysis. Despite the widespread use of internet-connected gadgets, connection problems persist due to the unavailability of the internet or limited network access. Edge analytics guards against possible network failures by ensuring that applications aren't hampered by internet issues or limited network access [17]. This is especially beneficial in rural regions (remote locations) or when trying to save communication expenses with costly technologies like cellular.

D. Deep Learning

Clouds are crucial in climate forecasting. Rainfall forecasting is also heavily influenced by the state and kinds of clouds in the sky. CNN model is used with transfer learning to classify clouds based on their features like texture, color, etc. to know the type of cloud and rainfall can be predicted using the precipitation associated with that cloud. Ground-based cloud images are readily available compared to satellite cloud images and provides information about the local atmosphere by analyzing bottom-level features of clouds like cloud height, cloud type, and cloud cover [18]. Rain clouds are mainly classified as Cirrus, Stratus, and Cumulus [19] and the main clouds are shown in Fig. 5.

The proposed device is provided with 8 megapixels digital camera integrated into Raspberry-pi to capture live sky images at any location by clicking a button on the device to activate the device. A cloud image is given as input to the convolution neural network (CNN) and the SoftMax activation function is used to classify the cloud into No Rain to very Low Rain, Low to Medium rain, or Medium to High Rain based on the amount of precipitation associated with each cloud. According to the Precipitation and the amount of Rainfall as shown in Table III, all cloud images are classified into three classes or groups.

The operation of the deep learning model as shown in Fig. 6 where the weights of the pretrained model are downloaded first, followed by freezing of all the layers except the top layer, which is used for classification in the fresh dataset, and finally the training of the model. All of the layers are unfrozen for fine-tuning, and the model is trained on a new dataset with a very low learning rate. Once trained, the model was able to accurately predict the outcome. All three pretrained models, VGG16, Inception-V3, and Xception, follow

the identical flow and concluded with The Xception model gives better accuracy to predict Rainfall based on cloud images taken from the ground, giving output whether Rain or little Rain or medium Rain or high Rain as compared to VGG 16 and Inception-V3. Rain Prediction using deep learning on ground-based cloud images using transfer learning, is presented in [20].

E. Machine Learning

To support the rain prediction model based on cloud images, atmospheric parameters are also considered which also plays vital role in determining the rainfall. Temperature, humidity, and pressure in the atmosphere fluctuate rapidly, causing instability in the atmosphere, which can result in rain, storms, and even lightning and thunder. The most significant elements in predicting precipitation are temperature and humidity [21]. Air pressure, dewpoint temperature (or relative humidity), wind speed, and cloud cover are four more meteorological factors that are substantially connected to rainfall [22]. But as the device is portable so only the most important variables are chosen for the experiments which are temperature, relative humidity, and air pressure that can be sensed with affordable sensors and are sufficient to predict the status of the rainfall.

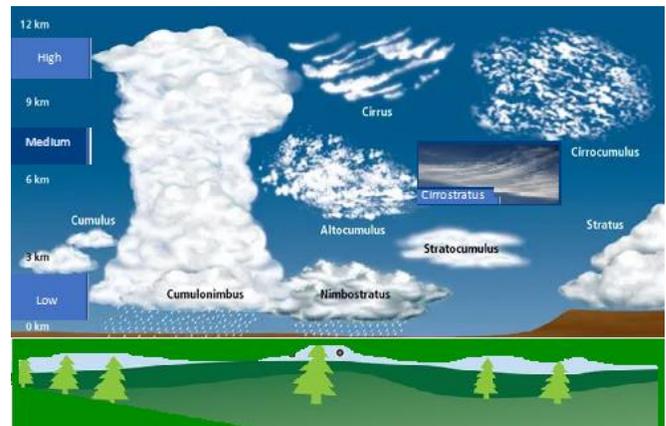


Fig. 5. Types of Clouds.

TABLE III. CLOUDS CLASSIFICATION BASED ON ASSOCIATED PRECIPITATION

Cloud	Associated Precipitation	Class
Cirrus Cirrostratus Cirrocumulus Altostratus	None None None None	No Rain to Very Low Rain
Stratus Stratocumulus Nimbostratus	Produces light showers or sprinkles, drizzle or Bring a light or moderate Rainfall of long duration	Low to Medium Rain
Cumulus Cumulonimbus	Showers or snow Heavy Rain with lightning, hail, or snow	Medium to Heavy Rain

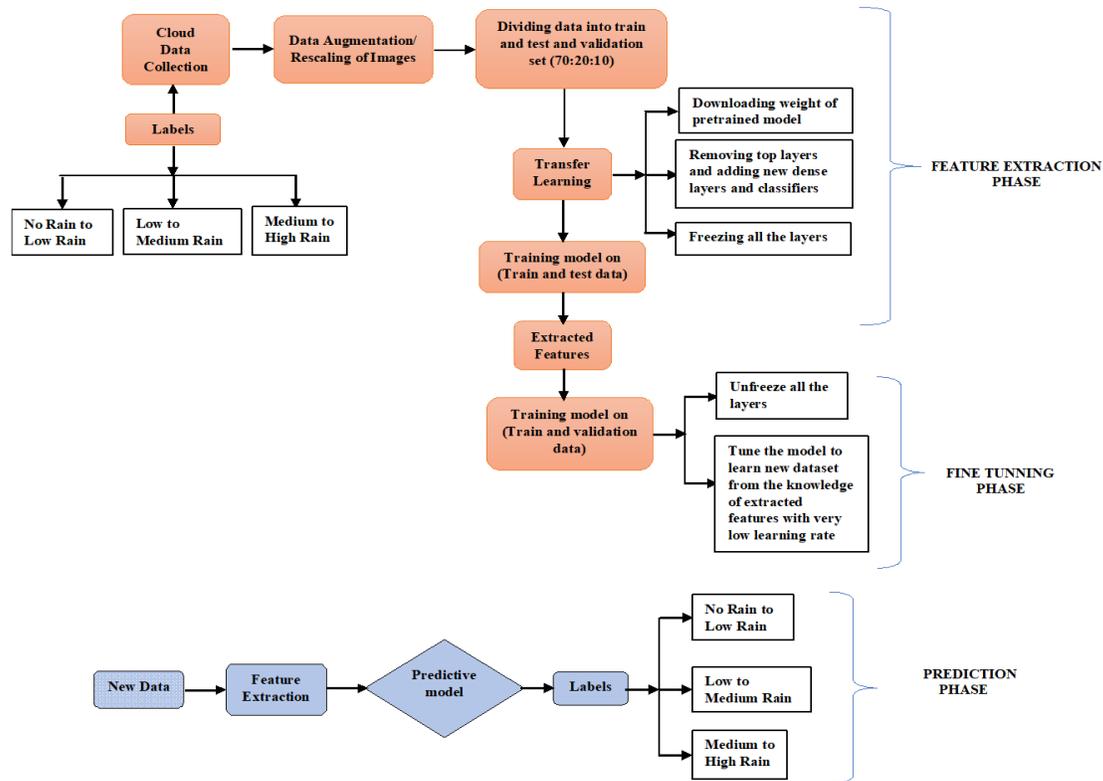


Fig. 6. Working of Deep Learning Model.

Rainfall intensity is classified according to the rate of precipitation, which depends on the considered time. The following categories are used to classify rainfall intensity [23]:

- Low rain — when the precipitation rate is < 2.5 mm (0.098 in) per hour
- Medium rain — when the precipitation rate is between 2.5 mm (0.098 in) and 7.6 mm (0.30 in) or 10 mm (0.39 in) per hour
- High rain — when the precipitation rate is > 7.6 mm (0.30 in) per hour, or between 10 mm (0.39 in) and 50 mm (2.0 in) per hour
- Very High rain — when the precipitation rate is > 50 mm (2.0 in) per hour

Here also the data is classified in the same categories as in cloud dataset as No Rain to Very Low Rain, Low to Medium Rain, and Medium to Heavy Rain based on the precipitation rate.

The data is collected for the city of Hyderabad, Telangana, from Nasa Power Data viewer for Daily data, which contains various climatic parameters, but only the most significant ones are picked, and the data is collected over a 40-year period from 1981 to May 2021. Because there are an average of 64-65 rainy days each year, collecting sufficient data over a long period of time is required. The dataset contains many attributes (columns), but for this experiment, the independent variables are Temperature, Relative Humidity, and Pressure,

and the dependent variable is Condition, which is labelled in three categories based on precipitation: No rain to very low rain, Low to medium rain, and Medium to high rain. Various steps carried out to train the model from collecting dataset to cleaning of the dataset, balancing the dataset, feature extraction, and then partitioning of the dataset into train and test sets as shown in Fig. 7. After that various classification models are applied on the training set to train the model on selected parameters then the performance is analyzed by calculating the accuracy based on the predictions from the test dataset.

It has been observed, that among all individual machine learning models RandomForest and KNN gave a good prediction as compared to others. Logistic regression and SVM also performed well while predicting the values but Decision Tree and Naïve Byes performance was poor in estimating the predictions compared to other models as shown in Fig. 8. For evaluating the machine learning models a very effective technique called K-FOLD CROSS VALIDATION is used where the model is tested on part of the dataset which was not used for training. In this experiment, value of k is set to 5 in the K-fold cross-validation technique to check the accuracy attained by various models used. Although predictions from stacking Ensemble are better than voting, the calculation time is three times that of individual machine learning models. As a result, the RandomForest classifier was chosen to implement in hardware since it is faster and has similar accuracy to Stacking Ensemble.

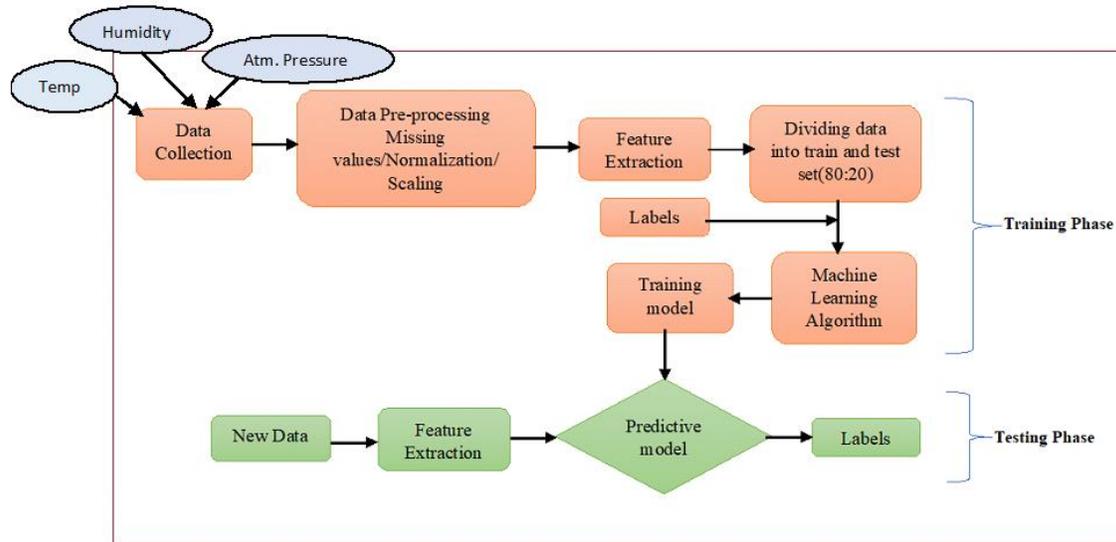


Fig. 7. Machine Learning Model for Rain Prediction based on Atmospheric Parameters.

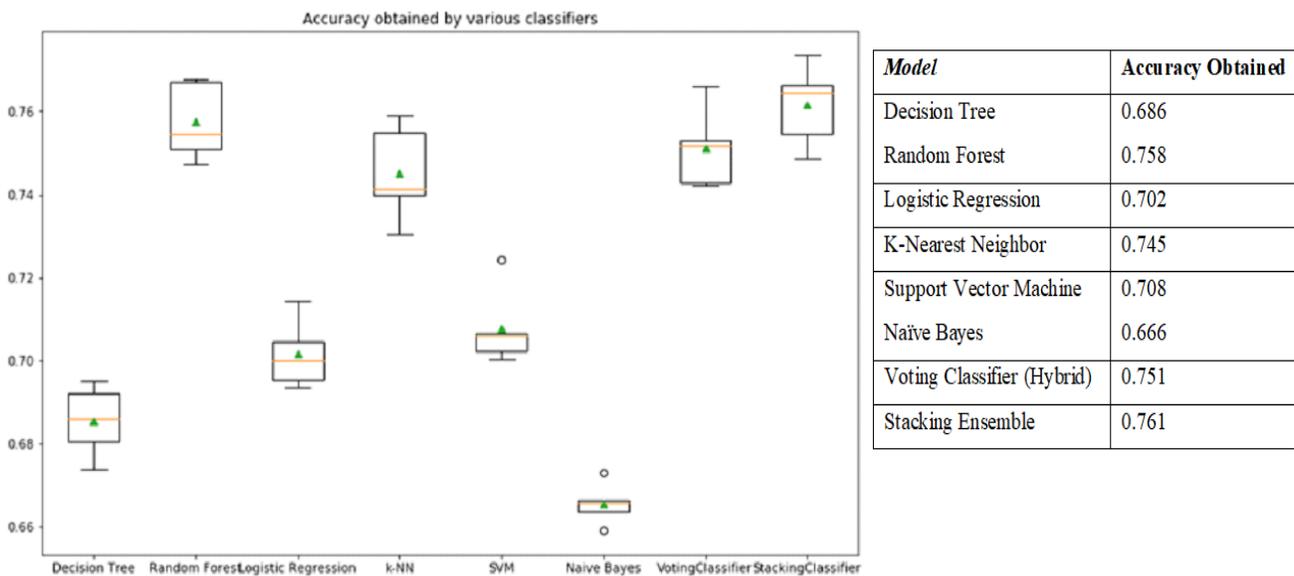


Fig. 8. Accuracy Obtained by Different Classifiers and Ensemble.

F. Device Setup

All the components are connected to Raspberry Pi as shown in Fig. 9. The screen displayed the currently sensed parameter values when the power is given to the device.

No internet or WIFI module is used in this device and the power can be given by the power bank to make the portable device so that it can be carried to any location. Logitech USB Camera is used to capture the video of the sky which will be divided into continuous image frames and given as input to the Deep learning model to get the rainfall prediction based on the clouds present in the sky. Data sensed by the DHT11 sensor and BMP180 i.e. temperature, humidity, and pressure is given as input to the machine learning model. The predicted rainfall from both the models is displayed on the OLED screen and

the final decision for the amount of irrigation is also displayed on the screen. Based on the output the motor can be switched ON to irrigate the land for the appropriate amount of water based on the decision given. The instrument is tested and found to give very correct predictions based on both approaches. This rainfall system would be very much useful in many agricultural events where the rainfall prediction needs to be considered. Prediction is given based on two different models where atmospheric parameters and clouds both are considered. The most important is this device works without the internet only power is needed. The device can also be used only to monitor individual atmospheric parameters temperature, humidity, or pressure which can be used to make a decision on food storage, or crop harvesting.

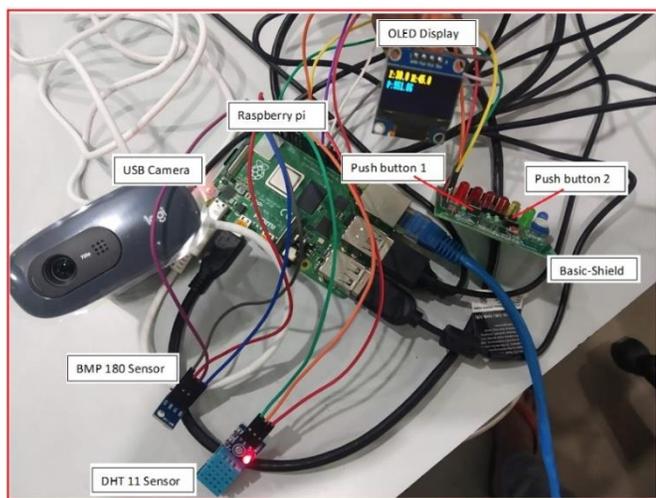


Fig. 9. Device Setup.

G. Tools and Libraries used

To execute the Deep learning model Debian GNU/Linux (64 bit) aarch 64(beta version) is installed which is a free and open-source operating. Debian GNU/Linux is a unique software distribution that combines Debian's philosophy and methodology with GNU tools, the Linux kernel, and other key free software to make aarch 64. Because of its technical brilliance and profound dedication to the requirements and aspirations of the Linux community, Debian is especially popular among expert users. The Deep Learning model and machine learning models were trained on python 3 so all the supporting libraries with the required versions Keras – 2.4.3,

Tensor Flow-2.4.0-rc2, Scikitlearn-0.20.2, and OpenCV -4.5.3 are installed.

V. RESULT

Some of the screenshots of output are shown in Fig. 10. Here in the display screen T denotes Temperature, H denotes Humidity and P denotes Pressure sensed by the sensor and are displayed as soon as Raspberry Pi starts. The Value of T, H, and P keeps on displaying on the screen. When button b1 is pressed its status changed from 1 to 0 and the rain prediction model starts by starting the camera to capture the sky image and is given as input to the deep learning and T, H, and P values are passed to the machine learning model. P1 denotes output from deep learning (i.e. from clouds) and P2 denotes output from machine learning (i.e. from Parameters).

The device is tested in the open space in residential area by powering it with the USB power bank as the Raspberry Pi requires a 5volt input voltage, which is provided via the USB type-C connection. The input voltage should actually be a little higher than 5 volts. Because power losses occur in the connectors and wires of the circuit's transmission, 5.1–5.2 Volts would be optimum. The results obtained as the intensity of rainfall were compared with the open weather API for the same day and time and found that the device prediction gives 70-75% accurate results which will be really helpful in many areas that require instant information on the rain at particular location and time. The device must be operated in open space and camera should be kept facing towards the sky to capture the cloud images. Device can be used as many times by rotating the facing of camera to capture sky in all directions.

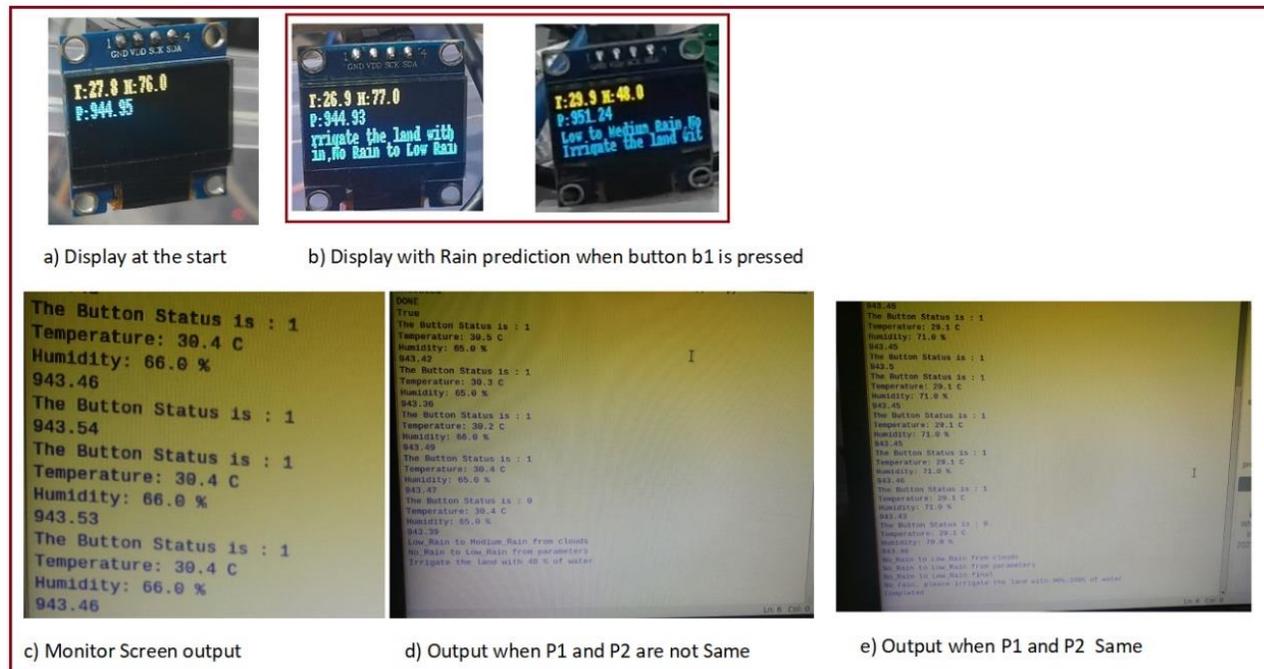


Fig. 10. Final Output at Different Stages.

VI. CONCLUSION AND FUTURE SCOPE

In this era of technology people like farmers must be able to use various technologies to increase their decision-making in various agriculture processes like the timing of irrigation, spraying of pesticides, use of fertilizers in the right amount, and at right time needed by the crops. But there are some limitations as unavailability of WIFI, No Internet, lack of knowledge while using the technologies. This IoT based portable device is a small effort for people like farmers, transporters where rain prediction is of utmost use for decision making and the main advantage of this device is that it will be operated without the internet and will give the prediction by sensing the atmospheric parameters and sky status at the current location. Compared to the previous devices this proposed system gives rainfall prediction based on atmospheric parameters along with present cloud/sky status specific to that location, hence is very useful in agriculture for monitoring the weather status without the use of any internet and the device require very low power which can be given by a power bank using USB-C Type cable to a raspberry pi.

Farmers cultivating crops under irrigation can benefit from climate prediction of precipitation and temperature at various stages of the growing season. These forecasts enable farmers to better manage the timing of water application and apply the appropriate amount of water to maximize crop production. This device can also be used to monitor individual parameters as per the requirement like temperature, humidity, or atmospheric pressure. The device works in two modes for monitoring individual parameters and gives rain prediction only after pressing the push button. Rain prediction is given by combining output from machine learning by giving atmospheric parameters as input and deep learning by taking sky image as input. For a deep learning model Xception model gives better accuracy of around 80% as compared to VGG16 and Inception V3. Whereas for the Machine learning approach among all individual machine learning models RandomForest and KNN gave a good prediction as compared to others. Logistic regression and SVM also performed well while predicting the values but Decision Tree and Naïve Byes performance was poor in estimating the predictions compared to other models. The device is handy and requires human intervention for pressing the button to get the rainfall prediction as per the requirement. Before reaching the final decision of irrigating the land the device can be used after every 1 or 2 hours to monitor the changes in the atmosphere. The sensed data can be stored in excel in Raspberry-pi and whenever a WIFI is available the data can be stored on the cloud for future use.

For future work, the GSM module can be integrated to operate the device automatically at regular intervals and the prediction can be sent to the farmer's mobile and also the motor can be switched ON/OFF as per the suggestion based on the rainfall prediction. Water wastage may be greatly reduced by including a smart irrigation system, which can reduce water consumption by 20%. The integration of smart technology, such as machine learning, IoT, the web, and the mobile framework, has been a major driver in achieving sustainable precision irrigation. Some of the study's findings

show that sustainable precision irrigation management can help farmers achieve food security and avoid water constraint.

REFERENCES

- [1] Amutha, D. (2013). Present Status of Indian Agriculture. Available at SSRN 2739231
- [2] Ramesh, K.V., Rakesh, V. and Prakasa Rao, E.V.S.(2020) Application of Big Data Analytics and Artificial Intelligence in Agronomic Research. Indian Journal of Agronomy, 65, 383-395.
- [3] Talaviya, T., Shah, D., Patel, N., Yagnik, H., & Shah, M.(2020). Implementation of artificial intelligence in agriculture for optimisation of irrigation and application of pesticides and herbicides. Artificial Intelligence in Agriculture, 4,58-73. <https://doi.org/10.1016/j.aiia.2020.04.002>.
- [4] Singh, R., Singh, H., & Raghubanshi, A. S. (2019).Challenges and opportunities for agricultural sustainability in changing climate scenarios: a perspective on Indian agriculture. Tropical Ecology, 60(2), 167-185. <https://doi.org/10.1007/s42965-019-00029-w>.
- [5] García, L., Parra, L., Jimenez, J. M., Lloret, J., & Lorenz, P. (2020). IoT-based smart irrigation systems: An overview on the recent trends on sensors and IoT systems For irrigation in precision agriculture. Sensors, 20(4), 1042. <https://doi.org/10.3390/s20041042>.
- [6] Susmitha, A., Alakananda, T., Apoorva, M. L., & Ramesh, T. K. (2017, August). Automated Irrigation System using Weather Prediction for Efficient Usage of Water Resources. In IOP Conference Series: Materials Science and Engineering (Vol. 225, No. 1, p. 012232).IOPPublishing. <https://doi.org/10.1088/1757-899x/225/1/012232>.
- [7] Choudhary, S., Gaurav, V., Singh, A., & Agarwal, S. (2019). Autonomous crop irrigation system using artificial intelligence. Int. J. Eng. Adv. Technol, 8, 46-51.
- [8] Nigusie, E., Olwal, T., Musumba, G., Tegegne, T.,Lemma, A., & Mekuria, F. (2020). IoT-based irrigation management for smallholder farmers in rural sub-Saharan Africa. Procedia Computer Science, 177, 86-93. <https://doi.org/10.1016/j.procs.2020.10.015>.
- [9] Linker, R., Sylaios, G., Tsakmakis, I., Ramos, T., Simionesei, L., Plauborg, F., & Battilani, A. (2018). Sub-optimal model- based deficit irrigation scheduling with realistic weather forecasts. Irrigation Science, 36(6), 349-362. <https://doi.org/10.1007/s00271-018-0592-x>.
- [10] RL, R., & Umamageswari, A. (2018). Modern Irrigation based on Web Weather Forecast.
- [11] Abhyankar, V., Singh, A. G., Paul, P., Mehta, A., & Vidhya, S. (2019, March). Portable Autonomous Rain Prediction Model Using Machine Learning Algorithm. In 2019 International Conference on Vision Towards Emerging Trends in Communication and Networking (ViTECoN) (pp.1-4).IEEE. <https://doi.org/10.1109/vitecon.2019.8899704>.
- [12] Kondaveti, R., Reddy, A., & Palabtl, S. (2019,March).Smart Irrigation System Using Machine Learning and IOT. In 2019 International Conference on Vision Towards Emerging Trends in Communication and Networking (ViTECoN) (pp.1-11). IEEE. <https://doi.org/10.1109/vitecon.2019.8899433>.
- [13] Tsukahara, J., Fujimoto, Y., & Fudeyasu, H. (2019,July). Rainfall Forecasting by using Residual Network with Cloud Image and Humidity. In 2019 IEEE 17th International Conference on Industrial Informatics (INDIN) (Vol. 1, pp. 331- 336). IEEE. <https://doi.org/10.1109/indin41052.2019.8972197>.
- [14] Sadhukhan, M., Dasgupta, S., & Bhattacharya, I.(2021,May). An Intelligent Weather Prediction System Based on IOT. In 2021 Devices for Integrated Circuit (DevIC) (pp.528-532). IEEE. <https://doi.org/10.1109/devic50843.2021.9455883>.
- [15] Dukes, M. D., Zotarelli, L., & Morgan, K. T. (2010). Use of Irrigation Technologies for Vegetable Crops in Florida.Horttechnology,20,133-142. <https://doi.org/10.21273/horttech.20.1.133>.
- [16] Da Silva, A. L. B. R., Coolong, T., & Diaz-Perez, J. C. (2019). Principles of irrigation scheduling for vegetable crops in Georgia. University of Georgia Cooperative Extension Bulletin, 1511.

- [17] Jin, H., Jia, L., & Zhou, Z. (2020). Boosting Edge intelligence with collaborative cross- edge analytics. *IEEE Internet of Things journal*, 8(4), 2444- 2458. doi:10.1109/JIOT.2020.3034891.
- [18] Ye, L., Cao, Z., & Xiao, Y. (2017). DeepCloud:Ground- based cloud image categorization using deep convolutional features. *IEEE Transactions on Geoscience and Remote Sensing*, 55(10), 5729-5740. <https://doi.org/10.1109/tgrs.2017.2712809>.
- [19] Houze Jr, R. A. (2014). Types of Clouds in Earth's Atmosphere. In *International Geophysics* (Vol. 104, pp.3-23). Academic Press. <https://doi.org/10.1016/b978-0-12-374266-7.00001-9>.
- [20] Ambildhuke, G. M., & Banik, B. G. (2021). Transfer Learning Approach-An Efficient Method to Predict Rainfall Based on Ground- Based Cloud Images. *Ingénierie des Systèmes d'Information*,26(4). <https://doi.org/10.18280/isi.260402>.
- [21] Holley, D. M., Dorling, S. R., Steele, C. J., & Earl, N. (2014). A climatology of convective available Potential energy in Great Britain. *International Journal of Climatology*, 34(14), 3811-3824. <https://doi.org/10.1002/joc.3976>.
- [22] Lekouch, I., Lekouch, K., Muselli, M., Mongruel, A., Kabbachi, B., & Beysens, D. (2012). Rooftop dew, fog and rain collection in southwest Morocco and predictive dew modeling using neural networks. *Journal of Hydrology*, 448, 60- 72. <https://doi.org/10.1016/j.jhydrol.2012.04.004>.
- [23] Narvekar, M., & Fargose, P. (2015). Daily Weather Forecasting using Artificial Neural Network. *International Journal of Computer Applications*, 121(22), 9-13. <https://doi.org/10.5120/21830-5088>.

E-Evaluation based on CSE-UCLA Model Refers to Glickman Pattern for Evaluating the Leadership Training Program

Ketut Rusmulyani^{1*}

Doctorate Program Student on Educational Science
Universitas Pendidikan Ganesha, Bali, Indonesia

I Made Yudana², I Nyoman Natajaya³, Dewa Gede
Hendra Divayana⁴

Universitas Pendidikan Ganesha, Bali, Indonesia

Abstract—This study aimed to describe the implementation of the level III leadership training program at the human resource development agency. The evaluation process used the CSE-UCLA model that was divided into: (1) components of system assessment, (2) program planning, (3) program implementation, (4) program improvement, and (5) program certification. This study involved 100 participants from the human resource development agency as institutional leaders, heads of divisions and heads of sub-sectors, lecturers/Widyaiswara, implementers/committees, superiors of alumni/mentors, and leadership training participants. Data was collected through questionnaires, interviews, observation, and documentation. The data were analyzed by quantitative descriptive analysis and verified by the Glickman quadrant, while the weaknesses found in the evaluation used qualitative descriptive analysis. The results showed that the level of effectiveness in terms of system assessment component included good criteria with a percentage of 82.4%, program planning component included good criteria (86.4%), program implementation component included good criteria (82.8%), and program improvement component included good criteria (83.2%). Lastly, the program certification component included good criteria (83%). The implementation of this Leadership Training Program is a strategy in developing SCA competencies, both managerial competencies, technical competencies, and socio-cultural competencies, to create a world-class bureaucracy in 2025 through independent learning and learning through coaching and mentoring.

Keywords—E-evaluation; leadership training; evaluation of educational programs; CSE-UCLA; human resource development agency

I. INTRODUCTION

Leadership is a performance that is carried out by someone in a group in order to move and influence others in developing human resources through education[1]. Leaders must have good decision-making skills in carrying out responsibilities for all processes, from procurement to evaluating work programs and the workforce[2]. However, in reality, there are still institutional leaders who are not fully able to manage themselves to change their mindset according to leadership concepts and competencies[3]. The facts show that the quality of the relationship between leaders and subordinates is still less harmonious and relevant[4],[5]. Some of the results of previous studies show an effect of increasing employee performance by transformational leadership, but some cases of

efforts to get transformational leaders are less open and competitive [6].

One of the essential efforts to continuously improve the quality of human resources for the apparatus is through education and training. Efforts to develop human resources are dealing with the facts on the ground that the implementation of education and training experiences several problems and obstacles[7]. The training's less than optimal implementation is due to the quality of the staff, facilities, and curriculum substance[8]. This problem is far from the ideal condition that the implementation of leadership training produces training participants' outputs who can demonstrate adaptive leadership competence (adaptive leadership) in developing innovations[9],[10]. In addition, the implementation and development of competency certification in education and training has not yet reached the target and has not been optimal, so the provision of innovation in the implementation of education and training does not continue and even stops [11],[12].

The determination of the innovation program needs to be regulated in the short, medium, and long term with specific, measurable, achievable, realistic time limit criteria[13],[14]. The implemented programs in accordance with the provisions need to be evaluated for the process and post-training. Evaluation applied in education and training named program evaluation. Evaluation of education and training programs is an activity carried out by an evaluator to collect and analyze complete and accurate information about the object/program/service/particular policy being studied [15],[16]. The education and training evaluation results can be used as recommendations in making decisions[17]. The implementation of education and training is able to ensure the formation of leader characters who have operational insight to build good governance. The evaluation results of the implementation of the training show that the old pattern was limited to equipping participants with the competencies needed to become visionary leaders [18],[19]. The problem of cost is also becoming an issue in implementing education and training. Therefore, post-training evaluation can be a policy solution regarding comparing the results and benefits of training with the costs incurred [20]. The post-training evaluation that have been carried out so far have not had a comprehensive normative standard and involve all stakeholders, so the recommendations obtained from the

*Corresponding Author

evaluation of the education and training carried out have not been optimal. In addition, the strengths and weaknesses in the implementation of education and training are not studied in depth [21].

The introduction of this study has explained the problems and constraints as part of a development and coaching system implemented by the education and training institution. The implementation of education and training must be implemented and developed in-depth to equip the state apparatus with good leadership [22]. This study examines scientifically through the CSE-UCLA (Center for the Study of Evaluation-University of California in Los Angeles) model. The CSE-UCLA model was chosen because it is very suitable and has advantages in the form of program implementation stages that can introduce the existence of the program being evaluated and the sustainability of the impact of an educational program[23]. The suitability of expectations and optimal implementation of the model will be known through a follow-up evaluation process [24]. This research is evaluative, which aims to determine the effectiveness of the leadership training program in terms of system assessment, program planning, program implementation, program improvement, and program certification in implementing programs in human resource institutions.

This study will observe the internal aspects of the education and training organization and its relation to implementing the education and training program that still needs to be improved. This study will observe the internal aspects of the education and training organization and its relation to implementing the training program, which still needs to be improved according to the objectives. The purpose of training in this research is to increase knowledge, skills, and attitudes in order to be able to carry out the duties of the level III leadership positions professionally based on the personality and ethics of civil servants according to the agency's needs. The fundamental objective of education and training in this research is through operational policies, which stipulate that the human resource development agency is a regional apparatus that carries out the functions of supporting government affairs in education and training. The training program's success is known through post-training evaluations to measure the level of success of the training process in an objective, reliable and valid manner, which is carried out after the training process is complete. The training program evaluation in this study used an e-evaluation based on the CSE-UCLA model, referring to the Glickman pattern for evaluating the leadership training program at the human resources development agency.

The research questions are (1) how is the effectiveness of implementing the leadership training program in terms of the system assessment aspect, namely the legal basis, the organization's vision and mission, regulations for the implementation of leadership training, local government support, and stakeholder support?; (2) how is the effectiveness of the implementation of the leadership training program in terms of program planning aspects, namely the management of the implementation of leadership training in terms of the ability of teachers/widyaswara, committee readiness, facilities and infrastructures, and budget?; (3) how is the effectiveness

of the implementation of the leadership training program in terms of program implementation aspects, namely the achievement of innovation, the factors driving and inhibiting alumni innovation, and the impact of innovation on the organization?; (4) how is the effectiveness of the implementation of the leadership training program in terms of the program improvement aspect, namely the participants' reactions to the abilities of the teachers/widyaiswara, committee, materials, schedules, and training programs, as well as the behavior of participants after returning to their work units?; (5) how is the effectiveness of implementing the leadership training program in terms of the program certification aspect, namely the adaptive leadership competence of alumni?

II. LITERATURE REVIEW

A. System Assessment of Leadership Training

The analysis of training needs is an ongoing process of collecting data to determine education and training needs[25],[26]. Education and training can be developed to help organizations achieve goals based on the results of a needs analysis so that they are the basis for program success[27]. This analysis begins with a training needs assessment (needs assessment) which aims to collect information on the training program's needs[28]. The results of the analysis of training needs are helpful as a basis for making a decision and providing solutions, instructions on what to do, how to implement, and what results are obtained[29]. The accuracy will significantly influence the leadership development program in preparing the curriculum, materials, methods, and learning evaluation systems that will be carried out[30]. The evaluation that will be carried out is expected to provide information about the value and benefits of the objectives to be achieved, design, implementation, and impact to help make decisions, accountability, and increase understanding of the existing phenomena[31].

B. Program Planning of Leadership Training

One of the main activities in implementing education and training is to design the program (designing and constructing the education and training). Design (design) is a planning process that describes a sequence of activities (systematics) regarding a program[32]. The design and construction of the education and training program are planning the sequence of activities for the education and training aspect, which is a unanimous unity of the program[33]. There are three essential elements in each education and training design that need to be considered to improve activities for each individual, namely: 1) purpose (what must be achieved); 2) method (how to achieve the goal); 3) format (under what circumstances the determination of the existing design is to be achieved)[34]. Planning is the initial activity of management functions. Planning is the most crucial stage of a management function, especially in dealing with a dynamically changing external environment[35]. Strategic planning is the first step that agencies must take in order to be able to respond to the demands of the local, national and global strategic environment. The realization of the strategic plan is carried out through the selection of targets and priority programs that must be implemented so that the vision and mission are in line

with the potential, opportunities, and constraints faced in efforts to increase performance accountability[36].

C. Program Implementation of Leadership Training

Leadership Education and Training is training that provides insight, knowledge, expertise, skills, attitudes, and behavior in the field of apparatus leadership to achieve leadership competency requirements in certain structural levels[37]. Leadership training is carried out to achieve the leadership competency requirements of government officials following the level of structural positions[38]. The competencies built-in leadership training are operational and tactical leadership competencies indicated by the ability to develop character and integrity behavior, develop activity plans, describe the agency's vision and mission, collaborate internally and externally, innovate, and optimize internal and external potentials. External to the organization[39]. These competencies can be achieved by designing a curriculum structure that includes five stages of learning, including 1) Diagnostic Stage of Organizational Change Needs; 2) Taking Ownership Stage; 3) Designing Change and Team Building Phase; 4) Leadership Laboratory Stage; and 5) Evaluation Phase[40].

D. Program Improvement of Leadership Training

Leadership behavior is an exciting study because it opens up great opportunities for everyone to become a leader. The study of the behavior and type of a leader for government organizations is growing and is supported by a government organizational model that is increasingly leading to a corporate organizational model[41]. The demand for leaders to constantly bring new things to the organization is hope for every individual[42]. Leaders of change who become the jargon in the new pattern of leadership training are not only an expression but in the new pattern of training. Each participant is expected to be able to present projects that can bring about change in the workplace[43]. Organizational leadership can be enriched through education and training, not solely on The grand man theory[42].

E. Program Certification of Leadership Training

Training is essentially aimed at developing human resource competencies. These competencies are developed through a conducive learning process during the training program[44]. The output of the training program is expected to support adaptive leadership competencies, the concept of innovation, and organizational performance[45]. Adaptive leadership means leadership that quickly adapts to changes and new circumstances[46]. The need for adaptive leaders is due to complex challenges and not enough operational improvisation to meet challenges[47]. Strategic organizational changes are needed to display satisfactory organizational performance[48]. The state civil apparatus must create work productivity to achieve public services that lead to good governance and clean governance[49],[50]. Improving service quality can be achieved by recognizing the conditions and challenges faced[51]. Resolution of organizational problems can be achieved with the principles of accountability and innovation[52].

F. CSE-UCLA Evaluation Model

Each education and training institution must have the competence or ability to build human resources for the apparatus, which is realized through implementing the education and training administration system by paying attention to quality, namely input, process, and output[53]. Sustainability of expenditure is the final result of the training not ending. Education and training institutions must monitor alumni performance in the form of follow-up evaluations (post-training) to determine the effectiveness of competencies in their work units[54]. CSE-UCLA is an evaluation model that has five evaluation aspects, namely system assessment, program planning, program implementation, program improvement, and program certification[55]. The CSE-UCLA model can evaluate service programs that help human life, such as educational learning programs, banks, cooperatives, e-government, e-learning[56]. A system assessment is an evaluation that provides information about the state or position of the system[57]. Program planning is an evaluation that helps select specific programs that may be successful in meeting program needs[58]. Program Implementation is an evaluation that provides whether the program has been introduced to specific groups as planned[59]. Program Improvement is an evaluation that allows the organization to achieve specific achievements[60]. A program certification provides information about the value or use of the program [61].

III. METHOD

The implementation of this research was carried out at the Human Resources Development Agency in the Province of Bali. This research is evaluative research that aims to determine the effectiveness of the leadership training program in terms of system assessment, program planning, program implementation, program improvement, and program certification in program implementation. An explanation of the sample, evaluation model design, data collection, and analysis can be presented in this section.

A. Sample

The distribution of the population in this study is in Table I.

TABLE I. DISTRIBUTION OF THE RESEARCH POPULATION

No.	Population Source	Total (Person)
1	Head of Institution	1
2	Head and Head of Subdivision	15
3	Alumni/Mentor Supervisor	34
4	Level III Leadership Training Alumni from 5 (five) batches	147
5	Lecturer/Widyaiswara	15
6	Committee	10
Total		222

The sampling technique used in this study is a purposive random sampling technique. The considerations used in this purposive sampling are (a) the training participants are in direct contact with the committee and education staff (widyaiswara) every day during the training; (b) In carrying out their duties, the training participants also interact with the committee, widyaiswara and mentors/superiors of the training participants; (c) After sampling, the samples with the following composition were obtained in Table II.

TABLE II. RESEARCH SAMPLE

No.	Population Source	Total (Person)
1	Head of Institution	1
2	Head and Head of Subdivision	15
3	Lecturer/Widyaiswara	15
4	Executor/committee	10
5	Alumni/Mentor Supervisor	25
6	Level III leadership training participants	34
Total		100

B. Evaluation Model Design

The variables involved in this program evaluation research are the effectiveness of the leadership training program implementation as measured by the components of the system assessment, program planning, program implementation, program improvement, and program certification to measure the adaptive leadership competencies of alumni. The component of the evaluation aspect is in Table III.

TABLE III. DESIGN OF THE CSE-UCLA MODEL ON EVALUATION OF THE IMPLEMENTATION OF THE LEADERSHIP TRAINING PROGRAM

Component	Evaluation Aspect
A(System Assessment)	1. Organizational Vision and Mission and Objectives of the Implementation of Leadership training (<i>Diklatpim</i>) 2. Legal basis 3. Support from local government and stakeholders
B(Program Planning)	Readiness/ability: 1. Lecturer/Widyaiswara 2. Committee 3. Infrastructures 4. Budget
C(Program Implementation)	1. Implementation of Leadership in implementing innovation 2. Factors driving and inhibiting alumni innovation
D(Program Improvement)	1. Participants' reactions to the ability of training staff/widyaiswara, committee, infrastructure 2. Behavior of participants after returning to the work unit
E(Program Certification)	1. Adaptive leadership competence 2. Impact of organizational innovation

C. Data Collection

The instrument used is a questionnaire. Correct and representative conclusions are generated from informations that were obtained correctly, validly, and reliably. Researchers used four methods to seek information from primary and secondary data sources: education and training managers (institutional leaders, heads of divisions, sub-sectors, and teaching staff/widyaiswara) and alumni of Leadership Training. Researchers use triangulation and reference materials. Through triangulation, researchers have checked the findings of the data by comparing it with various data sources and methods as well as time. The triangulation used in this research is source triangulation and method triangulation, and time triangulation. The data used in this study is primary data which is the answers of various research respondents, namely women and men, different ages, education levels, ranks, and positions. Data collection was carried out from May to September 2021. The questionnaires were distributed through the Google form considering that when the research was implemented PPKM level.4 COVID 19 in Bali Province (Governor Regulation No.12/2021).

The data obtained in this study is numerical data through a questionnaire compiled using a Likert Scale model to measure respondents' opinions, attitudes, and perceptions regarding the effectiveness of the implementation of Leadership Training. The scoring format using a Likert scale model is in Table IV.

Data collection through questionnaires is equipped with a grid of instruments to guide the making of questionnaires, namely a questionnaire for teachers/widyaiswara, questionnaires for participants/training alumni, interview grids (leader, head/head of sub field of mentor, widyaiswara, committee and training, and education alumni), observation grids and evaluation documentation of the leadership training program implementation. Technical data collection can be demonstrated by going through the flow or framework in Fig. 1.

The mechanisms for calculating content validity using the Gregory Formula are:

- 1) Assessment of the instrument per item by using a scale, a scale of 1 – 4.
- 2) The scale is grouped. For example, a score of 1-2 is grouped to be less relevant, a score of 3-4 is grouped to be very relevant.
- 3) The results of the experts are tabulated in the form of a matrix.

TABLE IV. SCORE FORMAT IN LIKERT SCALE

Positive Statement (+)		Negative Statement (-)	
Score.1	Strongly Disagree	Score.1	Strongly Agree
Score.2	Disagree	Score.2	Agree
Score.3	Neutral/Sufficiently Agree	Score.3	Neutral/Sufficiently Agree
Score.4	Agree	Score.4	Disagree
Score.5	Strongly Agree	Score.5	Strongly Disagree

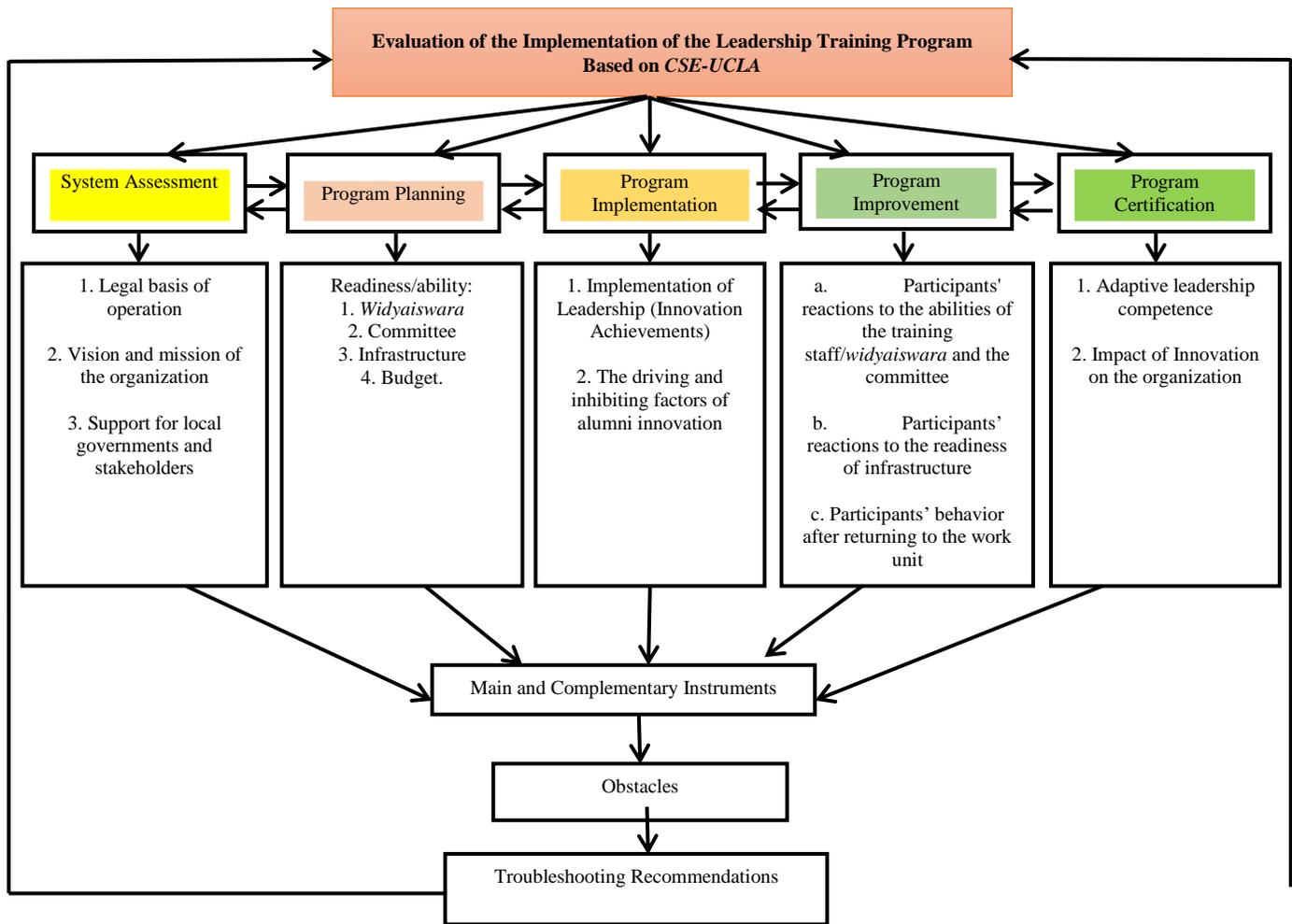


Fig. 1. Mechanism of Data Collection Evaluation of the Implementation of the CSE-UCLA-based Leadership Training Program.

4) Cross tabulation is made.

5) Content validity is calculated using the formula:

$$\text{Content validity} = \frac{D}{A+B+C+D} \quad (1)$$

Notes:

A = Disagreement between the two raters

B and C = Differences in views between raters

D = Valid agreement between the two raters

Content validity ≥ 0.60 is said to have good content validity and the test can be used in research.

For reliability, the formula used is Cronbach's Alpha as:

$$r_{11} = \left[\frac{k}{(k-1)} \right] \left(1 - \sum \frac{\sigma_b^2}{\sigma_t^2} \right) \quad (2)$$

Notes:

r_{11} = reliability,

k = number of questions,

2 = number of item variants,

t^2 = total variance.

Instrument reliability in its interpretation uses the interpretation of the correlation coefficients found to be large or small so that it can be guided by the provisions listed in the Table V [62].

By using this formula, the results of the empirical validity and reliability test are in Table VI. Based on the Table VI, it appears that the value of $r_{\text{count}} > r_{\text{critical}}$ for $n = 30 = 0.361$. Thus, all questionnaire items for the training and education alumni participants were declared valid and suitable to measure research variables. The questionnaire reliability coefficient for the training alumni participants was 0.949, with a very strong category based on the reliability calculation.

TABLE V. INSTRUMENTS' RELIABILITY

The value of r	Interpretation
0.00 – 0.199	Very low
0.20 – 0.399	Low
0.40 – 0.599	Moderate
0.60 – 0.799	Strong
0.80 – 1.000	Very strong

TABLE VI. VALIDITY AND RELIABILITY OF QUESTIONNAIRES OF THE TRAINING ALUMNI

Item Number	r _{count}	r _{critical} (n=30)	Conclusion	Reliability
1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43.	0.566; 0.571; 0.624; 0.381; 0.677; 0.626; 0.585; 0.605; 0.574; 0.659; 0.552; 0.600; 0.501; 0.589; 0.409; 0.546; 0.624; 0.381; 0.677; 0.534; 0.585; 0.605; 0.574; 0.659; 0.552; 0.614; 0.572; 0.531; 0.504; 0.749; 0.547; 0.562; 0.659; 0.529; 0.410; 0.520; 0.494; 0.481; 0.504; 0.545; 0.572; 0.562; 0.659.	0.361	Valid	0.949

Based on the Table VII, it appears that the value of $r_{count} > r_{critical}$ for $n = 30 = 0.361$. Thus, all questionnaire items for education and training teachers were declared valid and suitable to measure research variables. Based on the reliability calculation, the questionnaire reliability coefficient for education and training teachers is 0.938, with a very strong category.

TABLE VII. VALIDITY AND RELIABILITY OF THE EDUCATION AND TRAINING TEACHER QUESTIONNAIRES

Item Number	r _{count}	r _{critical} (n=30)	Conclusion	Reliability
1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45.	0.593; 0.530; 0.641; 0.503; 0.517; 0.478; 0.539; 0.641; 0.503; 0.430; 0.594; 0.527; 0.487; 0.612; 0.449; 0.663; 0.673; 0.424; 0.480; 0.497; 0.401; 0.646; 0.488; 0.405; 0.641; 0.503; 0.430; 0.594; 0.527; 0.487; 0.612; 0.449; 0.603; 0.673; 0.442; 0.394; 0.556; 0.521; 0.400; 0.401; 0.401; 0.646; 0.488; 0.552; 0.442.	0.361	Valid	0.938

D. Data Analysis Techniques

The data analysis techniques were carried out by

- 1) Distributing a questionnaire in terms of tangibles, reliability, responsiveness, assurance, and empathy;
- 2) Interviews with respondents to find out information about the leadership training program and synchronize the results of the questionnaires from the subjects studied, including the administrators of the educators themselves and the educators;
- 3) Observations are carried out to find out whether the implementation of the training seen from the observed aspects have been supported by authentic documents which are the legal basis for post-training implementation and to determine the effectiveness of the leadership training program in terms of aspects of system assessment, program planning, program implementation, program improvement, and program certification in the implementation of leadership training;
- 4) Documentation is related to the object of research and the program implementation.

Data analysis on the five CSE-UCLA Aspects was carried out quantitatively and qualitatively. Analysis of data on the evaluation of the implementation of the Leadership Training program in terms of the Aspects of the system assessment program planning, program implementation, program improvement, program certification using quantitative descriptive analysis measuring tools. For the constraints found in implementing this program using a qualitative descriptive analysis, primary and secondary data analysis was carried out in the data analysis stage. Primary data analysis includes:

Calculates the average score obtained in each evaluation and converts it into classifications and categories using a five scale based on the ideal mean (Mi) and ideal standard deviation (SDi), with the formula:

$$M_i = \frac{1}{2} (\text{ideal maximum value} + \text{ideal minimum value}) \quad (3)$$

$$SD_i = \frac{1}{6} (\text{ideal maximum value} - \text{ideal minimum value}) \quad (4)$$

The categorization of scores was determined by the three categories in Table VIII.

TABLE VIII. CATEGORIZATION OF SCORES IN THREE CATEGORIES USING IDEAL VALUES

Score categorization	Information
$(M_i) < \bar{X} \leq (M_i + 1.5SD_i)$	High
$(M_i - 1.5SD_i) < \bar{X} \leq (M_i)$	Enough
$\bar{X} \leq (M_i - 1.5SD_i)$	Less

Calculate the score obtained into the standard score (z score) with the formula.

$$z = \frac{x - \bar{X}}{SD} \quad (5)$$

z = Standard score

- X = raw score obtained by respondents
- \bar{X} = mean/mean
- SD = standard deviation/standard deviation

Change the standard Z-score into a T-Score with the formula:

$$T\text{-score} = (Z\text{-score} * 10) + 50 \quad (6)$$

Notes:

$T \geq 50$ Aspect value is high which is symbolized by H (Height)

$T < 50$ Aspect value is low which is symbolized by L (Low)

If 50 is a constant number which is the average limit of the normal curve that moves from 20 to 80 with six standard deviation values where one standard deviation of the value is 10.

Interpreting the T-Score of each Aspect into the Glickman Quadrant implementation level category is shown in Table IX.

TABLE IX. EFFECTIVENESS OF IMPLEMENTING A PROGRAM IN ALL ASPECTS OF CSE-UCLA FOLLOWING THE GLICKMAN PATTERN

Effective					Excellence				
A	B	C	D	E	A	B	C	D	E
H	H	H	H	L	H	H	H	H	H
H	H	H	L	H					
H	H	L	H	H					
H	L	H	H	H					
L	H	H	H	H					
Poor					Less				
A	B	C	D	E	A	B	C	D	E
L	L	L	L	L	H	L	L	L	L
					L	H	L	L	L
					L	L	H	L	L
					L	L	L	H	L
					L	L	L	L	H
					H	L	L	L	L
					L	L	H	L	L
					L	H	H	H	L
					L	H	L	H	H
					H	H	L	L	H
					L	H	L	H	L
					L	L	H	L	H
					H	L	H	L	L

To determine the effectiveness of a program or activity, the following classification is determined[63]:

Notes:

Excellence: If all five Aspects/variables are of high value (H)

Effective: If four of the five Aspects/variables are of high value (H)

Less : If one or two of the five Aspects/variables are of high value (H)

Poor : If all five Aspects/variables are low value (L)

Secondary data analysis includes: confirming the results of tabulation of primary data with data obtained through triangulation data collection techniques, namely interviews, observations, and cross-check documentation of the tabulation of the data obtained. They discussed and concluded things that resulted in an overview of the effectiveness of the implementation of the Level III Leadership Training program at BPSDM Bali Province in terms of the Aspects of the System Assessment, Program Planning, Program Implementation, Program Improvement, and Program Certification.

The data analyzed qualitatively include data obtained from interviews and observations or document studies. Furthermore, the validity of the data can also be checked. The data analysis technique, according to Sugiyono, follows the steps: data reduction, data display/data presentation, conclusion/verification[62]. Qualitative data validity test in qualitative research includes credibility, transferability, dependability, and confirmability tests.

IV. RESULT

Measurement of success is based on the interpretation of measurement behaviour that can describe the degree of quality and quantity and the existence of program implementation. However, the measurement results cannot be used as a reference for decision-making from quality and quantity if they do not have a comparison with a reference or comparison material. This research produces information about the effectiveness of the leadership training program in terms of system assessment, program planning, program implementation, program improvement, and program certification in program implementation.

A. Quantitative

Description of the quality of Leadership Training and Education data from the assessment of Teachers/Widyaiswara is shown in Table X. It can be concluded that it is univariate (analysis used on one variable to know and identify the characteristics of the variable). The tendency of scores in the description of the leadership training program is based on the data description and ideal score criteria. The determination of the ideal score criteria uses the ideal mean (Mi) and the ideal standard deviation (SDi) as a comparison to determine the score seen from the A-B-C-D-E (AP4) aspect. Table X, in general (average) the quality of the leadership training program on the AP4 aspect of the teacher/Widyaiswara assessment is good, namely 83.48%.

TABLE X. DESCRIPTION OF THE QUALITY SCORE OF LEADERSHIP TRAINING ON THE AP4 ASPECT OF THE TEACHER/WIDYAISWARA ASSESSMENT

The Calculation Results		Aspect					Implementation(AP4)
		A	B	C	D	E	
1	$\sum X$	61.87	64.9	62.11	62.5	62.25	312.96
2	N	15	15	15	15	15	15
3	M	4.12	4.32	4.14	4.16	4.15	-
4	SD	0.5	0.5	0.67	0.33	0.33	-
5	Mo	4.12	4.32	4.14	4.16	4.15	20.89
6	Mi	3.5	3.5	3	4	4	18
7	Mo (%)	82.4	86.4	82.8	83.2	83	83.48
8	Max	5	5	5	5	5	-
9	Min	2	2	1	3	3	-
10	Category	Good	Excellence	Excellence	Enough	Enough	

Data analysis using the Glickman Formula about the quality of the leadership training program obtained in this study was transformed into a T-score. Based on the data described on the quality of implementing the leadership training program for teachers/widyaiswara in terms of the CSE – UCLA Evaluation Model, it is summarized in the following Table XI. The quality of the Leadership Training Program Implementation for Teachers/Widyaiswara is categorized as Effective.

Quality results of implementation of leadership training programs for teachers/widyaiswara in view from the CSE-UCLA evaluation model using the Glickman formula can be seen visually in Fig. 2.

Description of leadership training quality data from the assessment of training alumni participants can be seen in Table XII. In general (on average), the quality of the leadership training program on the AP4 aspect of the assessment of the training alumni participants is very good, namely 90.72%.

TABLE XI. SUMMARY OF QUALITY RESULTS THE IMPLEMENTATION OF LEADERSHIP TRAINING PROGRAMS FOR TEACHERS/WIDYAISWARA IN VIEW FROM THE CSE-UCLA EVALUATION MODEL

Aspect		Frequency			Information
		High	Low	Results	
1	<i>A(System Assessment)</i>	5	10	Low	L
2	<i>B (Program Planning)</i>	8	7	High	H
3	<i>C(Program Implementation)</i>	8	7	High	H
4	<i>D(Program Improvement)</i>	8	7	High	H
5	<i>E(Program Certification)</i>	9	6	High	H

Notes:
L:Low
H:High

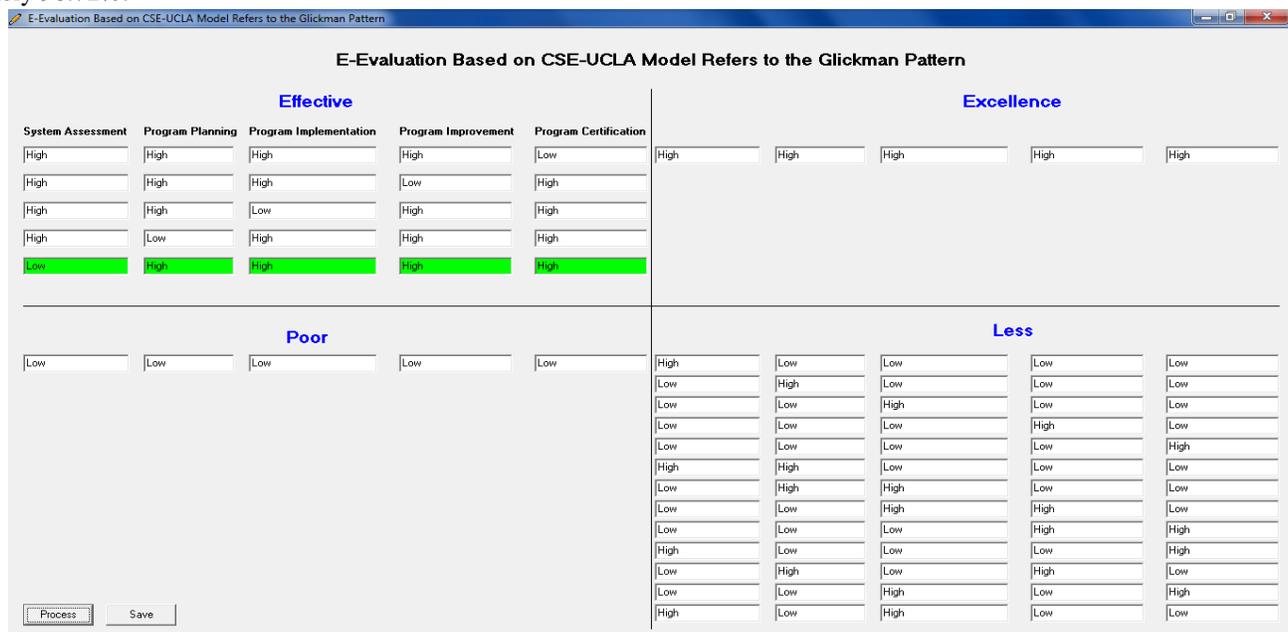


Fig. 2. Visualization of Glickman Formula Calculations in Determining Quality Results of Implementation of Leadership Training Programs for Teachers/Widyaiswara in View from the CSE-UCLA Evaluation Model.

TABLE XII. DESCRIPTION OF THE QUALITY SCORE OF THE IMPLEMENTATION OF THE LEADERSHIP TRAINING PROGRAM

The Calculation Results	Aspects					Implementation (AP4)
	A	B	C	D	E	
1 $\sum X$	229.4	224.13	221.76	229.8	228.4	1138.3
2 N	50	50	50	50	50	50
3 M	4.59	4.48	4.44	4.6	4.57	-
4 SD	0.5	0.5	0.67	0.33	0.33	-
5 Mo	4.59	4.48	4.44	4.6	4.57	4.57
6 Mi	3.5	3.5	3	4	4	18
7 Mo (%)	91.8	89.6	88.8	92	91.4	90.72%
8 Max	5	5	5	5	5	-
9 Min	2	2	1	3	3	-
10 Category	Excellence	Excellence	Excellence	Excellence	Excellence	

Based on the Table XIII, the quality of the leadership training program implementation for alumni training participants is categorized as Effective.

TABLE XIII. SUMMARY OF QUALITY RESULTS OF LEADERSHIP TRAINING PROGRAM IMPLEMENTATION FOR ALUMNI TRAINING PARTICIPANTS VIEWED FROM THE CSE-UCLA EVALUATION MODEL

Aspect	Frequency			Information
	High	Low	Results	
1 A(System Assessment)	32	18	High	H
2 B (Program Planning)	29	21	High	H
3 C(Program Implementation)	24	26	Low	L
4 D(Program Improvement)	35	15	High	H
5 E(Program Certification)	33	17	High	H

Notes:
L:Low
H:High

B. Qualitative

Description of Leadership Training quality data from the evaluation of the Committee, Teachers/Widyaiswara and alumni are presented in every aspect.

1) *Implementation effectiveness of leadership training program on the system assessment aspect:* I Nyoman Mariada said that the Human Resources Development Agency (BPSDM) of the Province of Bali as the center of excellence in developing the competence of SCA in the Province of Bali in supporting motto of Bali Government “Nangun Sat Kerthi Loka Bali” through a universally planned development pattern. The vision, mission and goals of BPSDM Bali support

the creation of the sanctity and harmony of Bali’s nature and its contents, to realize a prosperous and happy Balinese life with Trisakti Bung Karno principles (political sovereignty, economically dedicated, and culturally personal) through patterned development, comprehensive, planned, directed and integrated within the framework of the Unitary State of the Republic of Indonesia based on the values of Pancasila.

“The objective of *BPSDM* is to create professional and competent apparatus resources as stated in the strategic plan. The goal of this is to produce change leaders who are able to analyze and solve problems in their institutions in the form of innovations/change projects.”

Interview with Widyaiswara, I Made Sedana Yoga

“As a teacher, he is not directly involved in formulating and disseminating the vision, mission and objectives of the leadership training program implementation, but in general it is quite good. We strongly support this vision and mission and the programs it supports. So even though we are not directly involved in its formulation, we can see and judge, so far it has been a visionary.”

Interview with Widyaiswara, Dewa Ketut Winanda

Administratively, participants who wish to participate in leadership training at least have sat at the administrator level or supervisor level planned for promotion to administrator. They must also pass the selection and meet the portfolio and maximum age requirements. If this administrative selection passes, we can continue to participate and we will provide guidance according to the rules.”

Interview with Widyaiswara, I Made Gede Partha Kesuma

Administration standards are set as requirements and based on regulations from the center. The aim is to ensure that participants who take part in the training are individuals who have qualified and are able to become leaders for their respective institutions. Based on the results of interviews from several respondents from *widyaiswara* at *BPSDM* Bali, it can be seen that in organizing a program, an institution must have a clear vision, mission and goals.

2) *Implementation effectiveness of the leadership training program on the program planning aspect:* Widyaiswara’s readiness to teach is one of the keys to the success of implementing Leadership Training level III at the Bali Province BPSDM, so that the institution strives to continue to improve the competence and readiness of *widyaiswara* to be able to control and bring education and training towards the vision, mission and goals of the institution. As an effort to improve the competence of *widyaiswara*, BPSDM Bali Province must take an evaluation step to find out how the *widyaiswara*’s ability when teaching.

“For the purposes of improving the competence of *widyaiswara*, we routinely carry out performance evaluations during teaching. The main source on which we base our evaluation is using a questionnaire from the training participants. From the questionnaire, we will do a mapping in each aspect. Because from the questionnaire the answers are

usually mixed and subjective, we cannot immediately conclude from the results of the questionnaire, but also with other analyzes and considerations. Anyway, until later, we will generalize what needs to be evaluated.”

Interview with the Head of Sub Division of Core Technical Competency Development for Administration, Ni Ketut Suwardhyaksadewi

Facilities and infrastructure for the implementation of Level III Leadership Training at the Bali Province BPSDM have been well prepared before the program runs. So far, the Bali Province BPSDM has succeeded in facilitating all the needs of the Leadership Training and ensuring that both widyaiswara and training participants receive proper and functioning teaching and learning facilities.

“We can make sure that the infrastructure here is adequate and complete for the continuity of the education and training. And this is proven by the Pim III Education and Training having the A accreditation predicate, where one of the assessments to get A Accreditation is the availability of training infrastructure according to standards. So we are also grateful with all these facilities, we have succeeded in maintaining the quality standards of the education and training, we are also sure that from here the vision, mission and goals of the institution can be realized.”

Interview with Head of Sub Division of Learning Resources Management and Cooperation, Ni Putu Massuli Adi

3) *Implementation effectiveness of leadership training program on program implementation aspect:* The output of the implementation of leadership training is innovation in the form of change projects. The sustainability of the change (innovation) project for training participants is very important because there are still several milestones or goals in the change project that have not been achieved. This failure can usually be analyzed through the identification of the inhibiting factors.

“Actually, we already have innovation ideas for our institution. After participating in the training, we can also be strengthened about these innovations. However, conditions in the field sometimes hinder its realization, as in our opinion the most hindering is the hectic volume of work, even when on campus you still have to do office work, plus the number of staff is very minimal.”

Interview with Alumni of Training Participants, I Made Toya Arnawa

After the implementation of the Leadership Training, participants received reinforcement to implement change (innovation) projects at their respective home institutions. However, conditions in the field often create obstacles that complicate the realization of innovation in institutions. This is also similar to what was conveyed by other alumni of the training participants.

“After the training, we are enthusiastic to innovate, as if we have a lot of innovations that we can apply in institutions. We also get good ideas from mentors and widyaiswara. It’s

great if they are implemented. But indeed when we meet the facts on the ground, there are some obstacles. What we feel the most here is the lack of support, especially the support from the community.”

Interview with Alumni Training Participants, Ni Nyoman Sri Rahayu

The change (innovation) project is carried out in the institution under the guidance of a mentor. Mentors are direct supervisors of participants who have competence in providing support, guidance and input to participants to carry out change projects. During the formulation and implementation of the change project, the training participants must coordinate regularly with their respective mentors so that the innovations carried out are on target according to the initial design.

“The presence of a mentor here is very important, because sometimes we like to improvise when we innovate. Trying outside of the previous plan, so if it is not coordinated with the mentor, it may not be right on target. So we do continue to coordinate, discuss the concepts and technicalities, if there are obstacles, what are the solutions. Coordination with our mentor is quite good but the problem is that he has retired from duty but his successor is also very supportive until we can continue the realization of my change project.”

Interview with Alumni Training Participants, Ni Nyoman Sri Rahayu

4) *Implementation effectiveness of leadership training program on program improvement aspect:* Widyaiswara has been prepared by the institution to be able to teach the material in depth with a fun andragogy method. Mastery of the material from widyaiswara becomes one of the modalities of the successful implementation of Leadership Training. This widyaiswara readiness was also felt by other alumni of the training participants.

“There are some who have good quality both in terms of mastery of the material and the ability to distribute it to participants. Some may master the material, but for me it is not suitable in delivery so that it becomes less enthusiastic. But those whose names are teachers already understand that they have their own teaching styles. It may be suitable for me but not necessarily for others. But if you want to be objective in terms of mastery of the material, the widyaiswara at BPSDM have mastered it very well, so we can take a lot of knowledge.”

Interview with Alumni of Training Participants, I Made Toya Arnawa

The effectiveness of the implementation of the Level III Leadership Training program at BPSDM Bali Province in terms of the aspect of the program improvement can also be seen from the behavior of participants after returning to the work unit. Training can be said to be successful if it is able to provide positive changes to the training participants, both academically and non-academically.

“It is undeniable that we have gained a lot of knowledge from this leadership training. We share this knowledge with our colleagues at the institution, so that the benefits are not

only for us, but also for many people. On the other hand, what we get from the training we convey so that they understand and help carry out their duties responsibly. Because if we already know, then we can start and invite others. That's the principle."

Interview with Alumni Training Participants, Ni Wayan Vijayanthi Sari

Training participants gain new knowledge and experience to form problem-solving skills. Experience is gained mainly through change projects that bring participants face to face with various impediments to the implementation of ideas that need to be resolved. This is in accordance with the objectives of the innovation program (change project), where prospective change leaders (leadership training participants) at various levels are required to be able to identify and solve technical and adaptive problems.

"Personally, I feel that after participating in the leadership training, we know better how to deal with and overcome the problems that exist in the institution. After completing the training, the ability to solve problems faced in the work unit is more structured. We have the knowledge to be more observant to see what is behind the problem. Previously, we only knew that there was a problem, but we didn't really know what caused the problem, so it couldn't be solved. With this training, we can better understand how to deal with it all."

Interview with Alumni Training Participants, Ni Nyoman Sri Rahayu

5) *Implementation effectiveness of leadership training program on program certification aspect:* The results of this data analysis are also supported by interviews with alumni of the training participants regarding the participants' adaptive leadership competencies and the impact of organizational innovation after participating in the Leadership Training program. The main goal of Leadership Training is to produce visionary, innovative and creative leaders. These competencies are part of the character of adaptive leaders (adaptive leadership). Adaptive leadership after attending Leadership Training can be seen from the aspect of innovation replication and the ability to encourage the growth of a culture of innovation in the employees/staff of the training participants.

"Innovation replication that can be applied in work units by training participants will definitely be implemented. We learn a lot from mentors, as well as from other friends. There we observe the ideas that arise, if there are any that fit or approach the needs of the organization/work unit. Because innovation must look at the problems that exist in the organizational unit, so we don't just formulate the innovation. There are many considerations, and mentors are always accompanied."

Interview with Alumni Training Participants, Ni Nyoman Sri Rahayu

The increased performance of the Leadership Training participants is the result of an understanding of the main tasks and functions, as well as the inculcation of an attitude of

integrity from students. Leadership Training in addition to providing knowledge and experience, also fosters the fundamental values of a leader who has great responsibility in moving the wheels of the organization.

"The most important impact after participating in this leadership training is actually an increase in performance. Because performance is measurable in nature, before and after participating in the training, it can definitely be assessed if it is a performance problem. Frankly, our performance has increased, but whether it is significant or not, it goes back to the acceptance of each participant. But generally it increases. Because of that, we focus on receiving knowledge, we are given skills, experience. It all goes to the end of the performance improvement. Alumni performance is further improved because they get new knowledge during training that can be applied in work units."

Interview with Alumni Training Participants, Ni Nyoman Sri Rahayu

V. DISCUSSION

Based on data analyzed using system assessment aspects, program planning, program implementation, program improvement, and program certification, the implementation of the program from the teacher/widyaiswara assessment and alumni assessment is good. In the following, each aspect of AP4 is described and explained.

A. System Assessment

Aspects of the system assessment in Table X (assessment of teachers/widyaiswara) obtained 82.4% in the good category and the mean value of 4.12 in the good category. The data in Table XI (alumni assessment) scores 91.8% in the very good category, and the mean is 4.59 in the very good category. The concept of leadership is generally interpreted as a critical aspect in determining the success of an organization because leadership plays an essential role in the management of employees. The organization helps to maximize their efficiency (producing the most output with the minor input) and effectiveness to achieve their organizational goals[64]. The quality of the education and training program implementation is viewed from the vision and mission of the organization. The quality supported The objectives of implementing the education and training, the legal basis, and support from the local government and stakeholders. The vision, mission, and objectives of the leadership training program are following the direction of the institution's ideals and qualitative data analysis. Guided by a clear vision and mission and measurable goals, leadership training can run in a structured and systematic way for good development[28]. The widyaiswara interview on aspects of the system assessment shows that the compatibility between the vision, mission, and program objectives is structured based on the sanctity and harmony of nature, which creates a prosperous life. The formation of the vision, mission, and objectives is expected to produce good quality leaders and training programs. Therefore, the suitability of the leadership training's vision, mission, and objectives with the interview results of the training shows success for the further development of the training.

B. Program Planning

The aspect of program planning in Table X (assessment of teachers/widyaiswara) scores 86.4% in the very good category, and the mean is 4.32 in the very good category. For the data in Table XI (alumni assessment), the score is 89.6% in the very good category, and the mean is 4.48 in the very good category. Leadership training aims to develop apparatus resources regarding administrative skills, management (leadership), and leadership competencies[65]. The quality of the training program implementation is from the aspect of the planning program, the readiness/ability of the widyaiswara, the readiness/ability of the committee, the readiness of infrastructure, and budget readiness. The results of the qualitative data analysis show that the organizational structure of the Bali BPSDM has been structured and systematic. All teachers/widyaiswara have competence and quality in teaching. The facilities and infrastructure for implementing the Leadership Training of the Bali Province Human Resources Development Agency have been prepared well before the program runs. Level III Leadership Training and Education of the Bali Province Human Resources Development Agency also received an A accreditation rating, indicating that the implementation of the Level III Leadership Training Program, especially in the program planning component, has been running effectively[66]. The results of interviews with the head of the sub-sector on program planning indicate that forming an organizational structure in structured and systematic leadership training can help carry out organizational functions properly. Structured and systematic conditions make an enormous scope of work divided into small ones. The suitability of the institution's role with minor aspects will focus on performance. It can support the implementation of the vision, mission, and goals according to the aspect or field of expertise. This success and effectiveness make an excellent forum for implementing leadership training.

C. Program Implementation

Aspects of program implementation in Table X (assessment of teachers/widyaiswara) scored 82.8% in the very good category, and the mean value was 4.14 in the good category. For the data in Table XI (alumni assessment), the score is 88.8% in the very good category, and the mean is 4.44 in the very good category. The education and training program's quality support implementing innovation and the factors driving and inhibiting alumni innovation[67]. Qualitative data regarding the project also support the results of this data analysis. Changes (innovations) that can be implemented after attending the training Sustainability of innovation/alumni change projects can be used as a benchmark for improving agencies' performance in organizational units. The performance has succeeded in facilitating and guiding training participants to implement change projects at their respective home institutions. The realization of the change project encourages these new ideas to be realized[68]. The results of interviews with alumni of training participants on the program implementation aspect show that the implementation of sustainable innovation/projects is carried out in each institution because of the incompatibility of the education system, training implementation. This discrepancy is in implementing projects

that do not run optimally, such as doing office assignments during teaching activities on campus and coupled with minimal staff. The implementation of innovation development in each institution becomes an effective way. Implementing innovation development at each institution will not succeed optimally without guidance. In implementing this training, a mentor or mentor is presented from the participant's direct supervisor who has competence in providing support, guidance, and input in terms of innovation/project changes. The mentor system will support the effective implementation of the training following the expected goals.

D. Program Improvement

The aspect of the program improvement in Table X (assessment of teachers/widyaiswara) obtained a score of 83.2% in the good category and a mean value of 4.16 in the good category. For the data in Table XI (alumni assessment), the score is 92% in the very good category, and the mean is 4.6 in the very good category. The quality of the training program implementation gets from participants' reactions to the abilities of teachers/widyaiswara and committees, participants' reactions to the readiness of infrastructure, and participant behavior after returning to the work unit. Qualitative data analysis also shows that training participants gain new knowledge, experience, and skills to transform into adaptive leaders. This training program can support sustainability in developing new technologies and innovations[69]. These results indicate that the Level III Leadership Training of the Bali Province Human Resources Development Agency regarding the program improvement components has been running effectively. Aspects of program improvement can identify the causes of problems and formulate solutions to be said to be effective[70]. The results of interviews with alumni of the training participants on the aspect of the program improvement. The interview indicates that the accuracy of the objectives of the training and the process of developing the participants' leadership competencies is not always successful because the system has been structured. The aim of the training will show its success in the competencies possessed by each participant who has different thoughts and abilities. The training and education executive committee has prepared the program very well. The established program is to develop a coordination system for each participant to build good knowledge related to administrative and technical processes and infrastructure. Therefore, through a coordination system such as asking questions related to the development of education and training between participants, it makes it easy for participants to develop leadership competencies.

E. Program Certification

Aspects of the program certification in Table X (assessment of teachers/widyaiswara) obtained 83% with a fairly good category and a mean value of 4.15 in a good category. For the data in Table XI (alumni assessment), the score is 91.4% in the very good category, and the mean is 4.57 in the very good category. The quality of the implementation of the training program is viewed from the aspect of the adaptive leadership competency program certification and the impact of organizational innovation. The education and training evaluation results will be significant and influential

feedback in improving the quality of education and training and maintaining the sustainability of the education and training organization. The product of a training process is in the form of outputs or alumni of training participants, while further product benefits are in the form of outcome, namely how the influence of training on motivation which in turn has an impact on the actual performance of a training participant in its implementation[71]. Qualitative data analysis also shows a positive change in the behavior of the participants and an increase in the performance of the training so that in terms of the program certification component, the Level III Leadership Training and Development Agency of the Bali Province Human Resources Development Agency has been running effectively. This program can direct the leadership system to experience the development of organizational innovation[72]. The results of interviews with alumni of training participants on aspects of program improvement show that the program certification makes training participants have leadership competencies and has a good impact on organizational innovation. This success is the primary goal of implementing training that follows the character part of an adaptive leader. It can be said that the training participants who have carried out the training in a gradual and structured manner can encourage the growth of a culture of innovation in the employees/staff of the training participants. The education and training program, in general, has succeeded in encouraging training participants to play a role in cultivating innovation in their respective work units.

The results of research on all aspects of the system assessment, program planning, program implementation, program improvement, and program certification show that the implementation of leadership training can support the formation of the character of professional and integrity bureaucrats. Thus, training alumni can internalize the fundamental values of state civil apparatus SCA and the growth of high public ethics and do not stop at building competencies as has been going on so far but can produce leaders and change agents. The novelty of this program is that the development of Management knowledge with the CORPU (Corporate University) approach is a strategy in developing SCA competencies. Managerial competencies, technical competencies, and socio-cultural competencies to realize SMART SCA 2024 and world-class bureaucracy 2025 through independent learning and learning through coaching and mentoring. The CSE-UCLA model used in this study is considered to evaluate leadership training education programs[73]. The CSE-UCLA evaluation model has advantages that other evaluation models do not have, namely the program implementation stage, which can introduce the existence of the program being evaluated and the sustainability of the impact of an educational program[74]. This program can encourage training participants to innovate and is expected to transmit culture to innovate in their work environment.

VI. CONCLUSION

Based on the research findings and the results of data analysis, conclusions can be drawn about the effectiveness of the implementation of the leadership training program in terms of the components of the system assessment, program

planning, program implementation, program improvement, and program certification. The components in the entire system reach a minimum of good categories according to the assessment of the widyaswara and training alumni participants. The system assessment is running effectively with the support of the organization's vision and mission, the objectives of the education and training implementation, the legal basis, the support of the local government and stakeholders. Program planning is running effectively with structured and systematic support for widyaswara competencies, facilities, and infrastructure for the implementation of leadership training well prepared before the program runs. Program implementation runs effectively with the support of leadership implementation in carrying out innovations and the driving factors for the sustainability of alumni innovation. The program improvement is running effectively with the result that the training participants gain a lot of new knowledge, experience, and skills to transform into adaptive leadership in developing innovations.

VII. RECOMMENDATION

This research can provide recommendations for human resource development agencies to be responsible for preparing widyaiswara and leadership training program committees to have academic qualifications, competencies, and certifications to participate in the training program. Acceleration needs to be done to include managers who have not received MOT (Management of Training) training so that the quality of training management can be improved and the knowledge and skills of managers can be better. The purpose of organizing leadership training is to provide knowledge, skills, and attitudes in apparatus leadership. The human resource development agency needs to improve the development of knowledge management and the learning process by all organization members in identifying, creating, and distributing knowledge. The education and training program can be an alternative to encourage improvement in the implementation of the integrity zone and efforts to accelerate towards a corruption-free area. The CSE-UCLA model in this study has the advantage of measuring various aspects of the effectiveness of achieving the goals of the Level III Leadership Training Program. CSE-UCLA is related to the success of the decision-making process, and stages of the activities carried out. Furthermore, the CSE-UCLA evaluation model has advantages that other evaluation models do not have, namely the Implementation program stage, which can introduce the existence of the program being evaluated and the sustainability of the impact of a Leadership Training Program.

VIII. LIMITATIONS

In the context of management training, this research aims to evaluate the implementation and achievement of the goals and objectives of the Level III Leadership Training program. This research is needed in order to improve the quality of the implementation of the Level III Leadership Training program at BPSDM Bali Province. Implementation of evaluation through a series of interrelated processes to collect, analyze and report data is the focus of this research. The scope of the research is limited to research on the results of the implementation of the Level III leadership training program

with the sub-focus setting on five Aspects of the CSE-UCLA model at the Level III Leadership Training in 2019. The implementation of the CSE-UCLA Model in this training has limited program success, namely widyaiswara as a measure of success. The evaluator's task is more challenging, and he must be sensitive and have much dialogue. The evaluator became a living instrument before the evaluation criteria and tools were developed.

REFERENCES

- [1] M. Audenaert, A. Decramer, and B. George, "How to foster employee quality of life: The role of employee performance management and authentic leadership," *Eval. Program Plann.*, vol. 85, no. 101909, Apr. 2021, doi: <https://doi.org/10.1016/j.evalproplan.2021.101909>.
- [2] M. D. M. Alonso-Almeida and J. Llach, "Socially responsible companies: Are they the best workplace for millennials? A cross-national analysis," *Corp. Soc. Responsib. Environ. Manag.*, vol. 26, no. 1, pp. 238–247, Feb. 2019, doi: <https://doi.org/10.1002/csr.1675>.
- [3] P. E. D. Love, L. Ika, H. Luo, Y. Zhou, B. Zhong, and W. Fang, "Rework, failures, and unsafe behavior: Moving toward an error management mindset in construction," *IEEE Trans. Eng. Manag.*, pp. 1–13, Mei 2020, doi: [10.1109/TEM.2020.2982463](https://doi.org/10.1109/TEM.2020.2982463).
- [4] C. M. Barnes, E. Awtrey, L. Lucianetti, and G. Spreitzer, "Leader sleep deprivation, employee sleep, and unethical behavior," *Sleep Health*, vol. 6, no. 3, pp. 411–417, Jun. 2020, doi: <https://doi.org/10.1016/j.sleh.2019.12.001>.
- [5] B. Harb, B. Hachem, and H. Hamdan, "Public servants' perception of leadership style and its impact on organizational commitment," *Probl. Perspect. Manag.*, vol. 18, no. 4, pp. 319–333, Dec. 2020, doi: [10.21511/ppm.18\(4\).2020.26](https://doi.org/10.21511/ppm.18(4).2020.26).
- [6] F. Donkor, W. A. Appienti, and E. Achiaah, "The impact of transformational leadership style on employee turnover intention in state-owned enterprises in Ghana. The mediating role of organisational commitment," *Public Organiz Rev* 22, pp. 1–17, Feb. 2021, doi: <https://doi.org/10.1007/s11115-021-00509-5>.
- [7] R. Skiba, "Water industry cyber security human resources and training needs," *Int. J. Eng. Manag.*, vol. 4, no. 1, p. 11, 2020, doi: [10.11648/j.ijem.20200401.12](https://doi.org/10.11648/j.ijem.20200401.12).
- [8] M. Napal Fraile, A. Peñalva-Vélez, and A. Mendióroz Lacambra, "Development of Digital Competence in Secondary Education Teachers' Training," *Educ. Sci.*, vol. 8, no. 3, p. 104, Jul. 2018, doi: [10.3390/educsci8030104](https://doi.org/10.3390/educsci8030104).
- [9] M. I. Fadillah, "Competency development of trainers through professional partnership: an action research study for professional development," *MADIKA Media Inf. Dan Komun. Diklat Kepustakawanan*, vol. 5, no. 2, p. 7, 2020, [Online]. Available: <https://ejournal.perpusnas.go.id/md/article/view/696>
- [10] F.-Y. Lai, H.-C. Tang, S.-C. Lu, Y.-C. Lee, and C.-C. Lin, "Transformational Leadership and Job Performance: The Mediating Role of Work Engagement," *SAGE Open*, vol. 10, no. 1, p. 215824401989908, Jan. 2020, doi: [10.1177/2158244019899085](https://doi.org/10.1177/2158244019899085).
- [11] Stepanova, G. A., Tashcheva, A. I., Stepanova, O. P., Menshikov, P. V., Kassymova, G. K., Arpentieva, M. R., & Tokar, O. V., "The problem of management and implementation of innovative models of network interaction in inclusive education of persons with disabilities," *International journal of education and information technologies*. vol. 12, pp. 156-162, 2018, ISSN, 2074-1316.
- [12] S. Marr, K. McKibbin, A. Patel, J. M. Wilson, and L. M. Hillier, "The geriatric certificate program: collaborative partnerships for building capacity for a competent workforce," *Gerontol. Geriatr. Educ.*, vol. 42, no. 1, pp. 13–23, 2021, doi: <https://doi.org/10.1080/02701960.2019.1572004>.
- [13] J. I. Sentosa, "[The effectiveness of the implementation of the innovation program for level III leadership education and training participants in the human resource development agency of South Sumatra Province] Efektivitas implementasi program inovasi peserta pendidikan dan pelatihan kepemimpinan tingkat III di badan pengembangan sumber daya manusia Provinsi Sumatera Selatan," *J. Sumber Daya Apar.*, vol. 1, no. 2, p. 11, Nov. 2019, [Online]. Available: https://elibrary.diklatsumsel.id/uploads/041220201730_170-245-1-PB.pdf
- [14] R. Vaquero-Cristóbal et al., "Influence of an educational innovation program and digitally supported tasks on psychological aspects, motivational climate, and academic performance," *Educ. Sci.*, vol. 11, no. 12, p. 821, Dec. 2021, doi: [10.3390/educsci11120821](https://doi.org/10.3390/educsci11120821).
- [15] D. G. H. Divayana, A. Adiarta, and I. G. Sudirtha, "Instruments development of Tri Kaya Parisudha-based countenance model in evaluating the Blended Learning," *Int. J. Eng. Pedagogy IJEP*, vol. 9, no. 5, p. 55, Nov. 2019, doi: [10.3991/ijep.v9i5.11055](https://doi.org/10.3991/ijep.v9i5.11055).
- [16] S. J. Miah, M. Miah, and J. Shen, "Editorial note: Learning management systems and big data technologies for higher education," *Educ. Inf. Technol.*, vol. 25, no. 2, pp. 725–730, Mar. 2020, doi: [10.1007/s10639-020-10129-z](https://doi.org/10.1007/s10639-020-10129-z).
- [17] B. Gross et al., "Crew resource management training in healthcare: a systematic review of intervention design, training conditions and evaluation," *BMJ Open*, vol. 9, no. 2, p. e025247, Feb. 2019, doi: [10.1136/bmjopen-2018-025247](https://doi.org/10.1136/bmjopen-2018-025247).
- [18] N. Coers, "Cultivating visionary leaders to transform our world," *J. Leadersh. Educ.*, vol. 17, no. 1, pp. 1–6, Jan. 2018, doi: [10.12806/V17/I1/C1](https://doi.org/10.12806/V17/I1/C1).
- [19] J. Mascareño, E. Rietzschel, and B. Wisse, "Envisioning innovation: Does visionary leadership engender team innovative performance through goal alignment?," *Creat. Innov. Manag.*, vol. 29, no. 1, pp. 33–48, Mar. 2020, doi: <https://doi.org/10.1111/caim.12341>.
- [20] T. Samuel, R. Azen, and N. Campbell-Kyureghyan, "Evaluation of learning outcomes through multiple choice pre- and post-training assessments," *J. Educ. Learn.*, vol. 8, no. 3, p. 122, May 2019, doi: [10.5539/jel.v8n3p122](https://doi.org/10.5539/jel.v8n3p122).
- [21] H. M. Stallman, "Online needs-based and strengths-focused suicide prevention training: Evaluation of Care•Collaborate• Connect," *Aust. Psychol.*, vol. 55, no. 3, pp. 220–229, 2020, doi: <https://doi.org/10.1111/ap.12419>.
- [22] S. M. Loscalzo, T. Seimears, N. D. Spector, T. C. Sectish, and T. J. Sandora, "Leadership training in pediatric residency programs: Identifying content, characterizing practice, and planning for the future," *Acad. Pediatr.*, vol. 21, no. 5, pp. 772–776, 2021, doi: <https://doi.org/10.1016/j.acap.2021.03.016>.
- [23] D. G. H. Divayana, A. Adiarta, and I. B. G. S. Abadi, "Development of cse-ucla evaluation model modified by using weighted product in order to optimize digital library services in higher education of computer in Bali," *J. Pendidik. Vokasi*, vol. 7, no. 3, p. 16, 2017, [Online]. Available: <https://journal.uny.ac.id/index.php/jpv/article/view/13370/10272>
- [24] R. Plummer, J. Blythe, G. G. Gurney, S. Witkowski, and D. Armitage, "Transdisciplinary partnerships for sustainability: an evaluation guide," *Sustain. Sci.*, Jan. 2022, doi: [10.1007/s11625-021-01074-y](https://doi.org/10.1007/s11625-021-01074-y).
- [25] S. K. Pathy, "Training needs analysis for non-gazetted police personnel: An empirical study of commissionerate police in Odisha," *Int. J. Bus. Econ. Manag.*, vol. 2, no. 1, p. 13, 2019, doi: <https://doi.org/10.31295/ijbem.v2n1.63>.
- [26] İ. Aydın, B. Toptaş, A. Kaysili, G. Tanriverdi, N. Güngören, and Ş. Topçu, "Professional development needs analysis of school administrators and teachers in Turkey," *Kastamonu Eğitim Derg.*, vol. 29, no. 2, pp. 428–441, Apr. 2021, doi: [10.24106/kefdergi.821505](https://doi.org/10.24106/kefdergi.821505).
- [27] A. Gaureanu, A. C. Bejinariu, C. Feniser, and G. A. Paraschiva, "The analysis of vocational training needs. the case of romanian industrial enterprises," p. 8, 2018, [Online]. Available: <http://www.toknowpress.net/ISBN/978-961-6914-23-9/papers/ML2018-115.pdf>
- [28] C. Boon, D. N. Den Hartog, and D. P. Lepak, "A systematic review of human resource management systems and their measurement," *J. Manag.*, vol. 45, no. 6, pp. 2498–2537, Jul. 2019, doi: [10.1177/0149206318818718](https://doi.org/10.1177/0149206318818718).
- [29] G. Makransky, S. Borre-Gude, and R. E. Mayer, "Motivational and cognitive benefits of training in immersive virtual reality based on multiple assessments," *J. Comput. Assist. Learn.*, vol. 35, no. 6, pp. 691–707, Dec. 2019, doi: <https://doi.org/10.1111/jcal.12375>.
- [30] C. Sims, A. Carter, and A. M. De Peralta, "Do servant, transformational,

- transactional, and passive avoidant leadership styles influence mentoring competencies for faculty? A study of a gender equity leadership development program," *Hum. Resour. Dev. Q.*, vol. 32, no. 1, pp. 55–75, Spring 2021, doi: <https://doi.org/10.1002/hrdq.21408>.
- [31] N. V. Zolotykh, A. V. Chernyaeva, and T. U. Shevchenko, "Network support for personnel training: evaluation component," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 483, p. 012018, Mar. 2019, doi: 10.1088/1757-899X/483/1/012018.
- [32] D. B. Munari, A. L. G. Nogueira, E. T. Sousa, L. C. M. Ribeiro, and R. Sherman, "Sucessão de lideranças: uma reflexão necessária para o futuro da enfermagem," *Rev. Eletrônica Enferm.*, vol. 21, Dec. 2019, doi: 10.5216/ree.v21.54787.
- [33] E. Tingle, A. Corrales, and M. L. Peters, "Leadership development programs: investing in school principals," *Educ. Stud.*, vol. 45, no. 1, 2019, doi: <https://doi.org/10.1080/03055698.2017.1382332>.
- [34] A. F. Alheet, A. A. Adwan, A. Y. Areiqat, Ahmad. M. A. Zamil, and M. A. Saleh, "The effect of leadership styles on employees' innovative work behavior," *Manag. Sci. Lett.*, vol. 11, no. 1, pp. 239–246, 2021, doi: 10.5267/j.msl.2020.8.010.
- [35] S. Tafvelin, H. Hasson, S. Holmström, and U. von Thiele Schwarz, "Are formal leaders the only ones benefitting from leadership training? A shared leadership perspective," *J. Leadersh. Organ. Stud.*, vol. 26, no. 1, pp. 32–43, Feb. 2019, doi: 10.1177/1548051818774552.
- [36] M. W. True, I. Folaron, J. A. Colburn, J. L. Wardian, J. S. Hawley-Molloy, and J. D. Hartzell, "Leadership training in graduate medical education: Time for a requirement?," *Mil. Med.*, p. usz140, Jun. 2019, doi: 10.1093/milmed/usz140.
- [37] S. Holt, A. Hall, and A. Gilley, "Essential components of leadership development programs," *J. Manag. Issues*, vol. 30, no. 2, pp. 214–229 (16 pages), Summer 2018, [Online]. Available: <https://www.jstor.org/stable/45176579>
- [38] M. E. Ward et al., "Using co-design to develop a collective leadership intervention for healthcare teams to improve safety culture," *Int. J. Environ. Res. Public Health*, vol. 15, no. 6, p. 1182, Jun. 2018, doi: 10.3390/ijerph15061182.
- [39] T.-Y. Chen, "Medical leadership: An important and required competency for medical students," *Tzu Chi Med. J.*, vol. 30, no. 2, p. 66, 2018, doi: 10.4103/tcmj.tcmj_26_18.
- [40] A. Z. T. Arifani, A. Y. Susanti, and M. R. Mahaputra, "Literature review factors affecting employee performance: competence, compensation and leadership," *Dinasti Int. J. Econ. Finance Account.*, vol. 1, no. 3, pp. 538–549, Aug. 2020, doi: 10.38035/djefa.v1i3.491.
- [41] A. Till, J. McKimm, and T. Swanwick, "The importance of leadership development in medical curricula: A UK perspective (Stars are aligning)," *J. Healthc. Leadersh.*, vol. Volume 12, pp. 19–25, Mar. 2020, doi: 10.2147/JHL.S210326.
- [42] S. Tafvelin and A. Stenling, "A self-determination theory perspective on transfer of leadership training: The role of leader motivation," *J. Leadersh. Organ. Stud.*, vol. 28, no. 1, pp. 60–75, Feb. 2021, doi: 10.1177/1548051820962504.
- [43] D. Rosch, "Examining the (Lack of) effects associated with leadership training participation in higher education," *J. Leadersh. Educ.*, vol. 17, no. 4, pp. 169–184, Oct. 2018, doi: 10.12806/V17/I4/R10.
- [44] M. G. Goldsby, E. A. Goldsby, C. B. Neck, C. P. Neck, and R. Mathews, "Self-leadership: A four decade review of the literature and trainings," *Adm. Sci.*, vol. 11, no. 1, p. 25, Mar. 2021, doi: 10.3390/admsci11010025.
- [45] J. B. Cassel, B. Bowman, M. Rogers, L. H. Spragens, D. E. Meier, and The Palliative Care Leadership Centers, "Palliative care leadership centers are key to the diffusion of palliative care innovation," *Health Aff. (Millwood)*, vol. 37, no. 2, pp. 231–239, Feb. 2018, doi: 10.1377/hlthaff.2017.1122.
- [46] N. Thomas, "A review of the John Maxwell certification program," vol. 12, no. 1, p. 7, 2018, [Online]. Available: <https://digitalcommons.andrews.edu/cgi/viewcontent.cgi?article=1413&context=jacl>
- [47] A. A. Foster et al., "Strengthening and institutionalizing the leadership and management role of frontline nurses to advance Universal Health Coverage in Zambia," *Glob. Health Sci. Pract.*, vol. 6, no. 4, pp. 736–746, Dec. 2018, doi: 10.9745/GHSP-D-18-00067.
- [48] L. R. G. Lachter and J. P. Ruland, "Enhancing leadership and relationships by implementing a peer mentoring program," *Aust. Occup. Ther. J.*, vol. 65, no. 4, pp. 276–284, Aug. 2018, doi: <https://doi.org/10.1111/1440-1630.12471>.
- [49] N. Ashraf, O. Bandiera, E. Davenport, and S. S. Lee, "Losing prosociality in the quest for talent? sorting, selection, and productivity in the delivery of public services," *Am. Econ. Rev.*, vol. 110, no. 5, pp. 1355–1394, May 2020, doi: 10.1257/aer.20180326.
- [50] G. Wise, C. Dickinson, T. Katan, and M. C. Gallegos, "Inclusive higher education governance: managing stakeholders, strategy, structure and function," *Stud. High. Educ.*, vol. 45, no. 2, 2020, doi: <https://doi.org/10.1080/03075079.2018.1525698>.
- [51] K. Sendawula, S. Nakyejwe Kimuli, J. Bananuka, and G. Najjemba Muganga, "Training, employee engagement and employee performance: Evidence from Uganda's health sector," *Cogent Bus. Manag.*, vol. 5, no. 1, p. 1470891, Jan. 2018, doi: 10.1080/23311975.2018.1470891.
- [52] I. D. Raji et al., "Closing the ai accountability gap: Defining an end-to-end framework for internal algorithmic auditing," p. 12, 2020.
- [53] M. N. Dudin and Y. S. Shishalova, "Development of effective education and training system in the context of the transition to international accreditation," *Eur. J. Contemp. Educ.*, vol. 8, no. 1, Mar. 2019, doi: 10.13187/ejced.2019.1.118.
- [54] D. Kragt and H. Guenter, "Why and when leadership training predicts effectiveness," *Leadersh. Organ. Dev. J.*, vol. 39, no. 3, pp. 406–418, Jan. 2018, doi: 10.1108/LODJ-11-2016-0298.
- [55] L. Naibaho, "Online learning evaluation during Covid-19 using CSE-UCLA evaluation model at English education department Universitas Kristen Indonesia," *Bp. Int. Res. Crit. Inst. BIRCI-J. Humanit. Soc. Sci.*, vol. 4, no. 2, pp. 1987–1997, Apr. 2021, doi: 10.33258/birci.v4i2.1887.
- [56] T. Karalis, "Planning and evaluation during educational disruption: lessons learned from Covid-19 pandemic for treatment of emergencies in education," May 2020, doi: 10.5281/ZENODO.3789022.
- [57] D. Cullinan, E. A. Barnett, A. Ratledge, R. Welbeck, C. Belfield, and A. Lopez-Salazar, "Toward better college course placement: A guide to launching a multiple measures assessment system," *Community Coll. Res. Cent. Teach. Coll. Columbia Univ. MRDC*, Sep. 2018, doi: <https://doi.org/10.7916/D8892PK8>.
- [58] W. Zhan, "WEAVEonline: An assessment and planning management system for improving student learning," in 2017 Gulf Southwest Annual Regional Conference Proceedings, Richardson, TX, Mar. 2020, p. 33811. doi: 10.18260/1-2-1153-33811.
- [59] K. Phung and E. Ogunshile, "An algorithm for implementing a minimal stream X-Machine model to test the correctness of a system," *Int. Conf. Softw. Eng. Res. Innov. CONISOFT*, 2020, doi: 10.1109/CONISOFT50191.2020.00023.
- [60] S. Weiner et al., "Evaluation of a patient-collected audio audit and feedback quality improvement program on clinician attention to patient life context and health care costs in the veterans affairs health care system," *JAMA Netw. Open*, vol. 3, no. 7, p. e209644, Jul. 2020, doi: 10.1001/jamanetworkopen.2020.9644.
- [61] M. Orłowski, "External wine education and certification for restaurant service staff: a mixed-methods evaluation of training effectiveness," *Int. Hosp. Rev.*, Jul. 2021, doi: 10.1108/IHR-03-2021-0023.
- [62] Sugiyono, *Statistik untuk penelitian*. Bandung: Alfabeta, 2017.
- [63] M. Tegeh and N. Jampel, *Metode penelitian pengembangan*. Singaraja: Univeritas Pendidikan Ganesha, 2017.
- [64] A. J. Kaluza, F. Weber, R. van Dick, and N. M. Junker, "When and how health-oriented leadership relates to employee well-being—The role of expectations, self-care, and LMX," *J. Appl. Soc. Psychol. Publ.*, vol. 51, pp. 404–424, 2021, doi: DOI: 10.1111/jasp.12744.
- [65] C. M. Simamora, "Evaluasi pasca diklat kepemimpinan tingkat III tahun 2018 pusdiklat perdagangan Kementerian Perdagangan," *Cendekia Niaga*, vol. 3, no. 1, pp. 60–70, Oct. 2019, doi: 10.52391/jcn.v3i1.462.
- [66] N. Baw and S. Preudhikulpradab, "A roadmap for future development of leadership competencies of abc non-profit organization, Thailand and Myanmar," *ABAC ODI J. Vis. Action Outcome*, vol. 8, no. 2, p. 18, 2021, [Online]. Available:

- <http://www.assumptionjournal.au.edu/index.php/odijournal/article/view/5339/2974>
- [67] L. A. Juckett, L. Bunck, and K. S. Thomas, "The older Americans Act 2020 reauthorization: Overcoming barriers to service and program implementation," *Public Policy Aging Rep.*, vol. 32, no. 1, pp. 25–30, 2022, doi: <https://doi.org/10.1093/ppar/prab032>.
- [68] S. Ivanov, M. Belhassan, and C. E. Mahone, "Why great leadership principles remain largely irrelevant to modern enterprises: a special case study of a small moroccan company," *Int. J. Organ. Innov.*, vol. 10, no. 3, pp. 95–100, Jan. 2018, [Online].
- [69] K. Dolenc, A. Šorgo, and M. Ploj, "Perspectives on Lessons From the COVID-19 Outbreak for Post-pandemic Higher Education: Continuance Intention Model of Forced Online Distance Teaching," *Eur. J. Educ. Res.*, vol. 11, no. 1, pp. 163–177, Jan. 2022, doi: [10.12973/euler.11.1.163](https://doi.org/10.12973/euler.11.1.163).
- [70] R. Kum, I. S. Seo, T. H. Kim, S. W. Hahn, and M. S. Kim, "The effects of creative teaching technique applied to nursing major curriculum on critical thinking disposition, problem solving process, and self leadership," *J. Korea Converg. Soc.*, vol. 10, no. 3, pp. 373–382, Mar. 2019, doi: [10.15207/JKCS.2019.10.3.373](https://doi.org/10.15207/JKCS.2019.10.3.373).
- [71] A. Bharwani, T. Kline, and M. Patterson, "Perceptions of effective leadership in a medical school context," *Can. Med. Educ. J.*, vol. 10, no. 3, pp. e101–106, Jul. 2019, doi: [10.36834/cmej.53370](https://doi.org/10.36834/cmej.53370).
- [72] R. B. Hull, D. Robertson, and M. Mortimer, "Wicked leadership competencies for sustainability professionals: Definition, pedagogy, and assessment," *Sustain. J. Rec.*, vol. 11, no. 4, pp. 171–177, Aug. 2018, doi: [10.1089/sus.2018.0008](https://doi.org/10.1089/sus.2018.0008).
- [73] H. Koo and C. Park, "Foundation of leadership in Asia: Leader characteristics and leadership styles review and research agenda," *Asia Pac J Manag.*, vol. 35, pp. 697–718, 2018, doi: <https://doi.org/10.1007/s10490-017-9548-6>.
- [74] I. M. D. S. Atmaja, D. G. H. Divayana, and K. Setemen, "The design of the cse-ucla evaluation model using topsis and ahp methods for optimizing digital library services in badung regency," *J. Phys. Conf. Ser.*, vol. 1810, no. 1, p. 012036, Mar. 2021, doi: [10.1088/1742-6596/1810/1/012036](https://doi.org/10.1088/1742-6596/1810/1/012036).

Rule-based Text Extraction for Multimodal Knowledge Graph

Idza Aisara Norabid, Fariza Fauzi
Faculty of Information Science and Technology
Universiti Kebangsaan Malaysia, Bangi, Selangor, Malaysia

Abstract—Textual information is widely integrated in visual tasks such as object/scene detection and image annotation. However, the textual information is not fully exploited, overlooking the wide background knowledge available for Web images. This work proposes a multimodal knowledge graph (KG) to represent the knowledge extracted from unstructured Web image surrounding text and to integrate the relationship between image and text entities. Existing multimodal KG works have mainly focused on advanced visual processes for extracting entities and relations from images, and only employed standard text processing techniques such as tokenization, stop word removal, and part-of-speech (POS) tagging to capture nouns only or basic subject-verb-object from text in the semantic enrichment process. Adversely, neglecting other rich information in the text. Thus, the proposed approach attempts to address this as an automatic relation extraction (RE) problem to extract all possible triples from the text information from simple to complex sentences, in constructing the multimodal KG which eventually can be used as a training seed for visual tasks. A linguistic analysis is performed on a set of Web news articles consisting of news images and their related text. The dependency relations and POS information obtained are used to formulate a set of domain-agnostic entity-relation extraction rules. A triple extractor incorporating these rules, is developed to extract the triples from a news articles dataset and construct the proposed MKG. The Precision and Recall metrics are used to evaluate the extractor's performance. The evaluation results show that the proposed approach can extract entities and relations in the dataset with the precision score of 0.90 and recall score of 0.60. While the results are promising, the extraction rules can still be improved to capture all the knowledge.

Keywords—Relation extraction; knowledge graph; multimodal knowledge graph; dependency relations; object/scene detection

I. INTRODUCTION

The web consists of valuable images with surrounding textual information (also referred to as contextual information) that have been used in various computer vision tasks such as object/scene detection, image annotation, image clustering, image understanding, etc. This information is in which the surrounding texts are related to the image, rich in high-level semantic concepts and contains both direct and indirect information about the image [1-2].

While contextual information has been long used to improve computer vision task performance, there is still opportunity for improvement. According to [3], there is a gap between textual and visual analysis for image understanding. Existing algorithms, such as deep neural networks, mainly

focus on specific features in the image itself, typically ignoring the extensive background knowledge of the real world [4] that can be found in the contextual information.

In [3], [5] and [6], both visual and text are merged in the analysis for object detection and image retrieval, but these works focused more on the visual part rather than the text. The descriptions used are obtained from expert, hence, the image-text correlation is very high, however annotations from experts have the disadvantage of being time consuming and also expensive.

Many textual descriptions contain extensive background knowledge especially in news articles and blogs. The news carries variety of domain such as sport, education, entertainment and more. This implies that the knowledge is not restricted to a single area, but rather covers a broad range of topics. Hollink et al. [7] also states that news articles have a natural relationship between text and images. Works of [7] and [8] have proven the importance of images to be in their natural textual context, where they created corpus from online news articles. Typically, a news article will include a headline, the article and an image (or more) that is relevant to the news. Also, the image comes with a description (i.e., caption), where the description serves as context for the image and provides knowledge about it. By simply looking at the image, one may not fully comprehend the scene depicted. The thorough explanation of the scene is given in the description as well as in the headline and article, as illustrated in Fig. 1. Compared to manual captions from MSCOCO dataset in Fig. 2, the texts in these news articles are rich with high-level knowledge (abstract rather than object-level semantics), relevant to the article images, and readily available.

Larkin Sentral undergoes 9-hour sanitisation operation



Fig. 1. An Image and its Headline and Caption from a News Article.

several motorcyclists driving down a street into an intersection.
people riding down the street on their bikes.
a line of many motorcycles driving down the road.
a line of people on motorcycles on a street
a group of bikers riding bikes down a street.



Fig. 2. A Sample Image and its Captions from MSCOCO.

If the background knowledge in the Web news articles can be acquired in a structured manner, then a knowledge base can be generated, which can then be applied in visual tasks. Pan et al. [9] worked on methods that can improve the task of question answering (QA) and similarly, Wu et al. [10] proposed method to improve performance for answering visual questions. Both studies prove that it is important to incorporate external knowledge into machines to propose how humans handle such tasks. However, this vast knowledge that relates to the images is in the unstructured textual form. Thus, the suitable technique for extracting knowledge in a more structured way from large amounts of text and images and for describing the diversity of entities and concepts that exist in the real-world is required.

Li et al. [11] propose knowledge graph to learn knowledge for social image understanding. Knowledge graphs (KG) have long been used to represent knowledge and real-world events as graphs with nodes (entities), edges (relations) and labels. A KG is a data structure capable of capturing real-world concepts/entities and their relationships from unstructured text, transforming them into a structured graph. It reveals relationships between entities found in the text. Two entities and the relation between them are called triple (i.e., the basic building block for KG).

Gong and Wang [12] have produced a KG in their work where the graph is a multimodal KG (i.e., a KG consisting of text and images). The method used still has room for improvement as it only captures nouns or noun phrases, ignoring other rich information such as the verbs or the adjectives that explain about the image, hence, not fully utilizing the vast image background knowledge that can be acquired from the unstructured text.

To build a KG from text, the task of relation extraction (RE) is applied to extract the relationship. The development of this relation extraction algorithm has several techniques. The two general categories are rule-based and machine learning-based [13]. Rule-based relation extraction is performed by using linguistic knowledge and domain knowledge to build pattern based on words, parts of speech (part-of-speech) or semantics in collaboration with domain experts and then the

relationship is extracted according to the set of rules. Several works have used this approach in achieving their objectives [14,15,16].

Next, machine learning-based relation extraction methods use large amounts of labeled data for training and have shown good results in some instances [13]. There are also several open-source information extraction systems that use this approach such as NELL [17] and OLLIE [18]. However, problems can arise, such as lack of training data and poor generalized performance. Thus, a rule-based relation extraction is the best option for extracting relationships as it does not require prior training data and the set rules will extract a more overall result.

In conclusion, the issues discovered are that unstructured background knowledge (text) needs to be organized in a structured way while still maintaining the natural image-text relationship found in news articles. Second, the use of simple text (nouns) in the construction of multimodal knowledge graphs causes the overlook of other rich information available from the text. Therefore, the aim of this work is to build a multimodal knowledge graph by using the images accompanied by textual information in news articles, as well as linguistic-based relation extraction techniques to overcome the stated problems.

This paper contributes to the development of a set of rules for extracting entities and relationships from text and image in news articles. The rules are based on linguistic analysis of simple to complex sentences, factoring in the image from the news article. Grammar dependency relationships are utilized as they are syntactic rather than semantic, and thus, not restricted to any domain, and relationship between words is preserved as well. In addition, two new rules are introduced to preserve the inherent relation between the news image and text. A triple extractor is implemented using these rules. The resultant extracted triples are then used to construct a multimodal KG that represents the news image and its background knowledge, where the main entities and their relationships are clearly illustrated.

This paper is divided into several sections. Section 2 highlights issues related to multimodal KG and current linguistic-based relation extraction techniques. Section 3 discusses in depth the proposed framework to build the multimodal KG automatically from a corpus of news articles. The framework consists of three main phases: Phase 1: Pre-process datasets, Phase 2: Extract entities and relationships, and Phase 3: Build a multimodal KG. The triple extractor which produced the multimodal KG that describes the relationship between texts and images is evaluated based on precision and recall metrics and the findings are shown and discussed in Section 4 and finally, Section 5 concludes this paper.

II. RELATED WORK

This section provides an overview of KG and multimodal KG, with a focus on how text are processed and used in building multimodal KG as well as relation extraction (RE) techniques specifically the rule-based approach and dependency relationships.

A. Multimodal Knowledge Graphs (KG)

A knowledge graph (KG) is a directed labelled graph consisting of nodes and edges. Nodes are real-world concepts and apart from text, images can also be used as nodes. Edge connects a pair of nodes and shows the relationship between nodes [19]. Nodes can also be known as entities and the edges that connect these two entities are known as relations or relationships. Two entities and the relation between them are referred to as a triple, which are the basic building block for KG. Many existing KGs have been built such as Google Knowledge Graph, DBPedia, Wordnet, ConceptNet. These existing graphs have been used in many real-world applications including computer vision tasks such as object detection, visual question answering tasks, image classification and more.

Wu et al. [20] have built a KG from text in news articles. Since this work focuses on summarize output, the resulting graph is a summary KG even though the original input is a long sentence. The essence of the sentence has been captured; however, it may not capture other information found in a long sentence. For example, given the following two sentences:

“Two more young black men join in the beating, which is caught on cameras.”

“Two men who are young black and join fight.”

The first sentence is the original long input sentence which is summarized to the second shorter input text. Triples from the summarized sentences are used to build the KG. Hence, in the example, the output does not capture the phrase “caught on cameras” which can also be the entity and relation for another triple. In another example, “Alice and Bob took the train to visit the zoo. They saw a baby giraffe, a lion, and a flock of colorful tropical birds.” is summarized to “Alice and Bob visited the zoo and saw animals and birds”. The words in bold in the original longer sentence are the text that are dropped and summed up as “animals and birds”. These nouns can be important entities in computer vision tasks and the adjectives of the entities can be used to describe them. To obtain the main gist of a piece of information, the summarized sentences are sufficient, and the additional information may seem trivial. However, this trivial information can be considered particularly useful in object detection tasks even if it is only some descriptive text for certain entities.

Gong et al. [21] have produced multimodal learning approach for information extraction in which entities involve not only text but can include images or audio and relationships that connect entities either within or across modalities. The authors state that multimodal information such as text, pictures or audio are usually interrelated and complement to each other. Text and image modalities are focused due to the high availability of information.

Likewise, Gong and Wang [12] propose a multimodal learning algorithm to integrate textual information into visual knowledge extraction. While they have linked both the visual and textual parts in their proposed multimodal information extraction method, only nouns and noun phrases are used to tag image (or object in the image) with the “has-tag” relationship linking the image (or image object) to the image tags (i.e. the nouns or noun phrases), leaving out other rich information

available from text, for example, in the sentence “A girl is playing with a sleeping dog in a room”, the bold text “playing” and “sleeping” which give the actions for the nouns “girl” and “dog”, respectively, are not captured.

Attribute is generally used to describe an object. According to [22], attributes allow to describe, compare and categorize objects easily. Researchers such as [23] and [24] have proven that with the addition of these attribute, there have been and improvement in the visual tasks. Hence, the present multimodal KGs can still be further improved by filling in more information that is available from text into multimodal KG.

B. Relation Extraction (RE)

As mentioned in Section 1, to build a KG from text, it is very important to understand the text before extracting the relationship. Thus, leading to the task of relation extraction (RE). RE is a major sub-task of information extraction [25] and is also utilized for the detection and classification of semantic relationships between entity pairs [26].

Among the techniques in RE, [14] is one of the works that used a rule-based approach. The authors used this approach for mapping a predicate of a triple to an identical predicate in a KG. However, the generated rules cannot cover all possible patterns in open domain because of the sparsity of unstructured text. Similarly, in [15] use a rule-based approach but with the addition of a similarity-based approach to achieve their objective. In which, the resulting rules are able to cover all possible patterns which result in more complete triples.

In [16] has conducted a study to build an open information extraction system for Indonesian language with rule-based approach. This author has proven that by only using rule-based still can formulate a generalized rule that can capture triples in a wide, open domain. Thus, this work is referenced in terms of the method used to extract relationships between entities. They use part of speech (POS) tagging such as noun, verb, etc. and dependency relationships to extract relationships. The authors concluded that the method used was to identify the relationship based on the VERB POS tag in the extraction of the single verbs. Moreover, the ADVMOD (adverbial modifier) dependency relationship was used alongside VERB POS to obtain a more complete relationship. Extracting entities for both subject and object produced a complete triple. However, the author does not consider the syntactic relationship that exists between the texts which describes a word i.e., an adjective to a noun. By taking in consideration this type of relation, most of the relationships that exist between texts will be captured.

Overall, the reviewed RE methods perform well for simple sentences but poorly for complex sentences. This study attempts to consider adverbial phrases with the extraction of verbs and prepositions in addition to the extraction of single verbs and utilizes the adjectival modifier (AMOD) dependency relation where this relationship describes the nature of an entity; therefore, leveraging on the available text resources.

C. Dependency Relationships

The dependency-based parser labels the relationship that is dependent on the key word in order to get a sense of the

predicate-argument relationship [27]. The task of the dependency parser is to take the input text and apply the proper set of dependency relationships to it [28]. A dependency parser helps to create a dependency tree, which is a tree model based on dependency relationships, by parsing words or sentences.

TABLE I. RELATIONS IN CLAUSE PREDICATE CATEGORY

Clause Predicate	Description
Nsubj	Nominal subject
Nsubjpass	Passive nominal
Csubj	Clausal subject
Csubjpass	Clausal passive subject
Dobj	Direct object
Iobj	Indirect object
Ccomp	Clausal complement
Xcomp	Open clausal complement

TABLE II. RELATIONS IN NOUN DEPENDENTS (MODIFIER) CATEGORY

Noun Dependents	Description
Amod	Adjectival modifier
Advmod	Adverbial modifier
Nmod	Nominal modifier
Nummod	Numeric modifier
Appos	Appositional modifier
Det	Determiner
Compound	Compound

Based on Universal Dependencies (UD) [29], there are 42 relationships that can be grouped into nine categories: (1) clausal predicates, (2) Non-core dependents of clausal predicates, (3) clausal dependents, (4) Noun dependents,

(5) Coordination, (6) Compounding and unanalyzed, (7) Case-marking, prepositions, possessive, (8) Loose joining relations dan (9) others. According to [27], frequently used relationships focused on only two of the nine UD categories. The two categories are clausal predicates and noun dependents (modifiers). Table I and Table II are the examples of the list of universal dependency relationships that have existed. Some of these relationships, mainly subject (nsubj, nsubjpass, csubj, csubjpass), object (dobj and iobj), modifiers (amod, advmod) and compound relations, will be investigated in the analysis process of defining the rules.

III. PROPOSED METHOD

In this section, a detailed explanation of the framework for building a multimodal KG is given. The framework for this study has three main phases. Phase 1: pre-process dataset, Phase 2: extract entities and relationships and Phase 3: build a multimodal KG. Fig. 3 shows the flow of how a multimodal KG is built from a news article.

This study contributes to the extraction of entities and relationships from real-world sentences found in news articles that can be simple short sentences up to long and complex sentences. Simple sentence is the article headline, mainly a short sentence that only have one verb per sentence while complex sentence is the caption which is a long sentence that can consist of multiple verbs in a sentence. This study aims to extract the relationship between two entities or known as triple from the text. The dependency relationship technique is applied to maintain the relationship of each word. Linguistic analysis is performed on each sentence to obtain grammatical dependencies, where a set of dependency relationships will be identified. This is an initial step to detect consistent pattern to formulate relation extraction rules. Once the rules set have been formulated, this set will be used to extract triples of the text. Finally, the extracted triple is used to construct a multimodal KG.

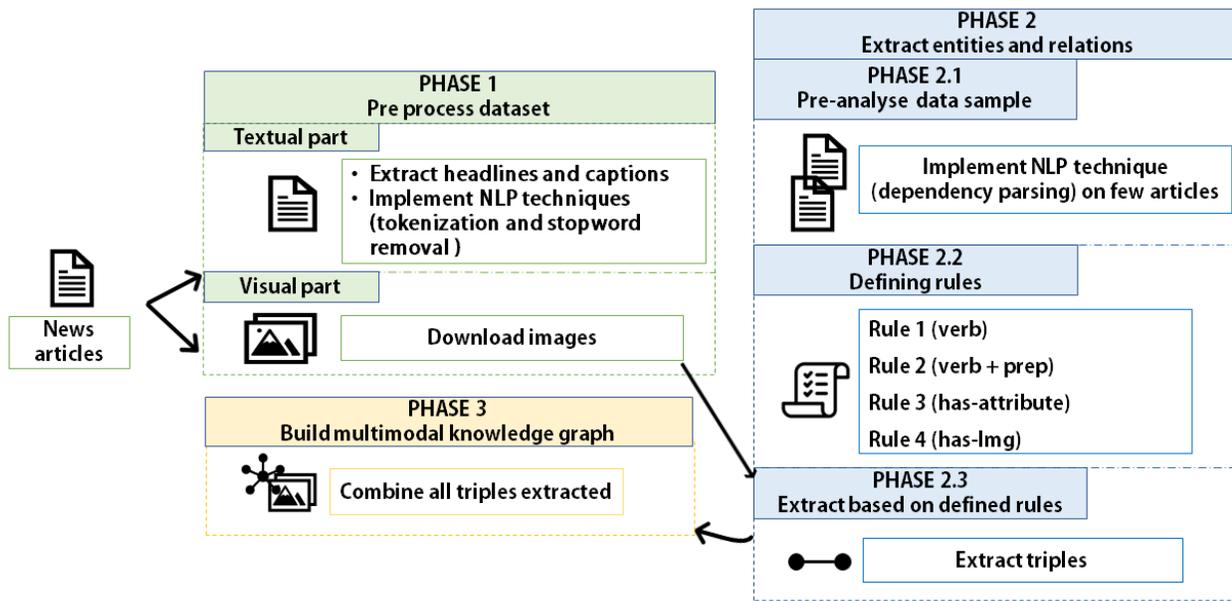


Fig. 3. Framework for Building a Multimodal KG.

A. Phase 1 (Pre-process Dataset)

Starting with the first phase, the Phase 1 input is a collection of online news articles. Each article consists of an image and a text article. Here, the focus is on the image and the descriptive text that accompanies the image. Therefore, this study only considers article titles, captions and images. The text source is derived from the title and caption of the image while the visual source is from the article image.

A web news article consisting of news text and images related to the news. News articles need to be cleaned and filtered because news articles have a lot of information gathered (headlines, captions, date published and more). However, in this study, only a small portion of this rich information will be used. This is because the title and caption are sufficient to describe the image. NLP techniques such as tokenization and stop word removal will be applied to titles and captions. Both techniques will be modified so that the output produced is consistent with the study.

To briefly described, tokens can be formed in individual words or phrases but usually, one word is detected as one token. Even so, in some cases, there are words that cannot be considered a token. Tokenization is customized in a way that can combine several words as a single token. This customization will collect all words that have a hyphen (-) and combine them into one token. The next pre-processing step is stopword removal. The original stopword list should not be used in this study. Words such as prepositions (at, in, of, etc.) can be used as examples of why the original list cannot be used. This is because words prepositional words will be used later to extract relationships. Thus, the stopword list will be self-determined to suit this study. Specifically, words like 'left, right, above, center, below'. These words are generally used in the caption to describe the position of a particular object in an image and are removed because it affects the performance of dependency-based parser which makes it less accurate.

For the visual part, each news article usually has an image to support the news. Each of these images will be downloaded via URL. It is then stored to be paired with the has-Image relationship which will be explained further in phase 2.2. This is to indicate that the caption is related to the image.

In conclusion, the output from this phase is the extracted text referring to the title and caption, and the image that has been downloaded. Finally, the dataset, D, represents the set of the entire news article.

$$D = \{T, V\} \quad (1)$$

where T and V are equivalent to the text and visual parts, respectively. For the set, T, consisting of h and c. h represents the title while c represents the caption.

$$\{h, c\} = T \quad (2)$$

For set V, which consists of only one element, img, which is the downloaded image.

$$\{img\} = V \quad (3)$$

B. Phase 2 (Dependency-based Entity Relation Extraction)

The next phase is an extension of the textual part which is extracting entities and relationships. It will be divided into three sub-phases namely Phase 2.1: pre-analyse data, Phase 2.2: defining rules and Phase 2.3: extract based on the defined rules. Briefly, in Phase 2.1, several articles consisting of simple and complex sentences were selected and parsed through a grammatical parser. In Phase 2.2, the output of the grammar parser which is the dependency tree of each article will be analysed. This is an initial step to detect consistent pattern to formulate relation extraction rules. Once the rules have been determined, the triples (entities and relationships) will be extracted in Phase 2.3. Also, the downloaded images earlier on will be used as one of the entities for the triple set. Typically, a sentence with only a single verb consists of one triple. However, in some cases, there can be one sentence consisting of several triples. For instance, long sentences that have multiple verbs will produce more than one set of triples.

1) *Phase 2.1 (pre-analyse data: dependency parsing analysis)*: In this sub-phase, the pre-analysis is carried out to identify rules for entity extraction. 10 sentences (headlines and captions) are randomly selected from several news articles together with the accompanying images and are analyzed manually which information supposed to be extracted. Then, it is parsed through a parser to show its grammatical structure (POS tag, dependency relations) to detect a pattern. The 10 sentences consist of five simple sentences and five complex sentences. Five more sentences are analyzed to ensure that no new patterns emerge. Hence, these 10 sentences are sufficient for the pre-analysis because of the nature of English sentences to have a similar grammar pattern, thus the result will be much alike.

Firstly, the set of sentences are manually examined where words that describe an image are identified and marked in yellow as shown in the Fig. 4 and 5. These marked words are considered as information that should be extracted. The sentences are put through a grammar dependency parser to obtain the parse tree structure, list of dependency relations and POS tag. Each word that has been marked is viewed in its grammatical structure that is the dependency relationship and POS tag as shown in Fig. 6. Based on the grammatical outputs produced, there are several dependency relationships that are often present on the marked words. The Visitation: Glasgow City Council pay family over Nazi-looted artwork. The crew of an Emirates Airline Boeing 777 prepares for passengers ahead of a demonstration flight in Dubai in 2007. Emirates announced plans Thursday to begin flying a Dubai to Panama City route on a 777, which will be the longest in the world.

Table III shows the frequently found dependency relations together with their explanations by [30,31].



Fig. 4. Example of Simple Sentence.



Fig. 5. Example of Complex Sentence.

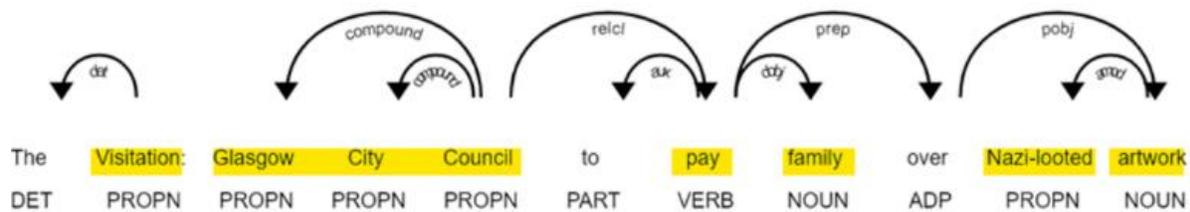


Fig. 6. Example of Dependency Tree.

TABLE III. TYPES OF DEPENDENCY RELATIONSHIPS

Dependency Relationships	Description
ROOT	the main topic (verb) for the sentence
Compound	nouns that modify the head of a noun phrase
Nsubj	nominal subject (noun phrase)
Dobj	direct object
Pobj	prepositional object
amod	adjective phrases that change the meaning of a noun phrase
Prep	prepositional phrases that modify the main meaning

According to [27], frequently used relationships are focused on two categories of relationships as mentioned in the previous section which is the clause predicate (Table I) and Noun Dependent (Table II). In Table III, these are the list of consistent relationships generated by the parser which fall into both of those relationship categories. The relationships are Compound, Nsubj, Dobj and amod. The prep relationship needs to be considered in order to capture the verb as a whole. Pobj in turn connects object entities to prep relationships.

According to [31], ROOT is a special label in the dependency tree that is usually on the main verb of a sentence. For some cases, if a phrase is processed or in other words is not a full sentence, ROOT is assigned to the noun of the head of the phrase. It can be observed that ROOT can be either a verb or a noun, in some sentences; and when the ROOT relation and POS noun are found in a sentence, the noun entity is highly related to the image, or to a particular entity in the image.

Hence, this is used to specify a new relation between the text (noun entity) and image (i.e., a text-visual relationship).

Apart from the dependency relationship is the POS tagging. This is the process of labelling punctuation in a language based on class classification. In other words, POS tagging indicate parts of speech to each word such as nouns, verbs, adjectives and more. Table IV shows the most common labels found on words that have been identified by the parser.

Having observed and analyzed the output of dependency relationships and POS tagging for these data samples, Tables III and IV are consistent patterns generated by the parser of the marked text. All these dependency relations and POS tag are used as the basis for the entity and relations extraction rules.

TABLE IV. POS TAG LABEL

POS tag	Description
ADP	Preposition
ADJ	Adjective
DET	Determiner
NOUN	Noun
PROPN	Proper nouns
VERB	Verb

Based on the output in Fig. 6, basically for a relationship between texts, the text to be captured as a relationship has a POS tag labelled VERB and the entity has a POS tag labelled NOUN or PROPN. But the proposed method will also use the dependency relationship that has been generated by the

dependency parser to extract the entities more accurately. The pattern for entity extraction based on dependency relationships is that text labeled Nsubj, Dobj or Pobj will be captured as an entity. The dependency relationship derived from the dependency tree also denotes a word that is either a subject or an object.

To extract the overall relation of the verb, the text having the 'prep' dependency relationship was also combined with the main verb. With this, making the extracted relationship was more ideal. Furthermore, texts describing an entity are extracted based on 'amod' dependency relationships making full use of all naturally occurring relationships between texts. In contrast to the text-image relationship, this relationship is pre-defined and uses the ROOT label to extract the appropriate entity.

Consistent pattern of dependency relationships and POS tag are used as the basis for rules for extracting entities and relationships. Also, relationships will be categorized into two namely 1) text-text and 2) text-visual. Category 1 is an extracted relationship that exists naturally between texts, for example, verb-based relationships, verb-based relationships and prepositions and relationships based on 'amod' dependency relationships but are named as has-Attribute while Category 2 is a relationship set to link text with images.

Once the entity and relation are successfully extracted completely, then the triple, R that is completely extracted will be produced and the relation (h), subject (s), and object (o) are arranged in the following order.

$$R = \{h, s, o\} \tag{4}$$

2) Phase 2.2 (Defining rules): As noted earlier, this study focuses on rule-based relation extraction and covers not only text-text relationships but also text-visual relationships. In this subphase, four types of rules are set for extracting relationships as shown in Fig. 7. Therefore, the development of the rules is explained in more detail. After the analysis in

Phase 2.1 is performed, the following rules are defined for extracting the relationship:

- Rule 1 (based on verb relationships).
- Rule 2 (based on verb + preposition relationship).
- Rule 3 (based on has-Attribute relationship).
- Rule 4 (based on has-Image relationship).

For Rule 1, the relation is captured first before the entity, so the first step is to identify the verb. The token having the ROOT dependency label and verb POS tag (VERB) will be captured as its relation. Next is to find subjects and objects as entities.

As for Rule 2, in addition to the verb itself, the relationship of verb + preposition is also considered in this study; similar to Rule 1, but with the addition of a prep dependency relationship as in Fig. 7.

Next, this Rule 3 is based on an amod-dependency relationship renamed as 'has-Attribute' relationship to indicate the characteristics of a particular word, which is mostly taken from an adjective word. Other than the previous rule, this Rule 3 will identify the object first by identifying the word with the amod dependence.

Since the has-Image relationship has been predefined so Rule 4 is defined to extract the entities for the relationship that link the image and text. Thus, it represents a text-visual relationship.

3) Phase 2.3 (Extract based on rules): Once the rules are determined, then the relations and entities will be extracted according to the rules as described in Phase 2.2. The processed data will go through an algorithm and a triple list will be generated. It will be arranged according to the respective articles. In this way it is clearer that there are some sentences that have more than one triple especially complex sentence.

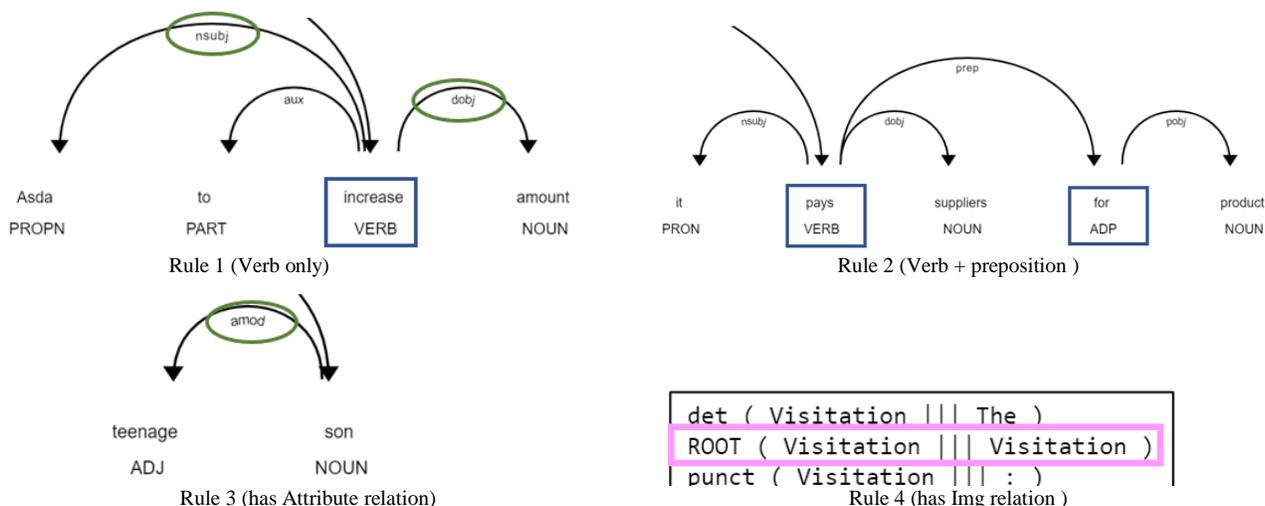


Fig. 7. List of Rules.

C. Phase 3 (Build a Multimodal KG)

In this phase, the relationship of texts and text-visual will be described. The extracted triples will be combined and produce a multimodal knowledge graph which can be formulated as follows. G represents a multimodal knowledge graph for an article that contains a combination of several subgraphs or triples, R_n where n is the total number of triples extracted.

$$G = \{R_1, R_2, \dots R_n\} \quad (5)$$

The multimodal KG can be visualized as in Fig. 8 where the graph has two different types of modalities namely text and image. From the graph, there are other subgraphs. A subgraph produced represents a triple.

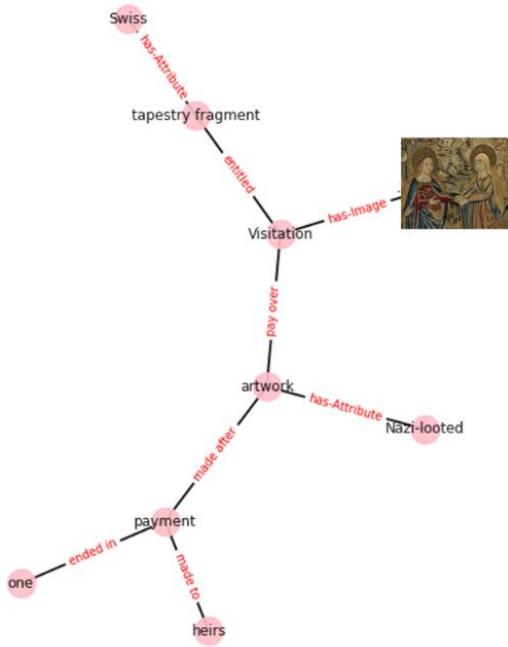


Fig. 8. Multimodal Knowledge Graph.

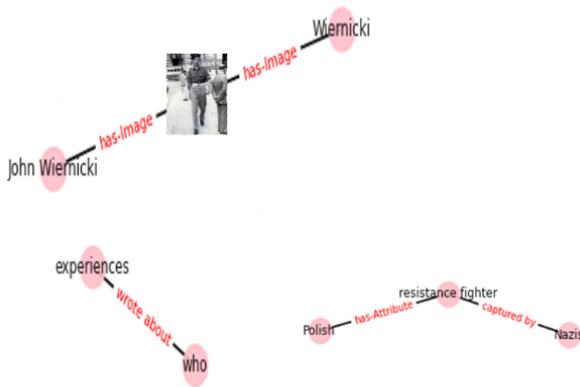


Fig. 9. Hanging Sub-graph.

However, after the experiment was done it was found that there was a subgraph that hung alone as in the following example of Fig. 9. For human-like mindset, it can be concluded that the hanging subgraph still has some relation with the image. This relationship can be labelled as an indirect relationship. Thus, the addition of the rule (Rule 5) for the indirect relations will be applied to the subgraph. This relationship will be called has-Bg-Kg where Bg represents Background while Kg represents Knowledge. This relationship indicates that the subgraph describes the background knowledge of an image in addition to creating a stronger relationship between other subgraphs and also the image-text relationship.

IV. EXPERIMENT AND DISCUSSION

A. Experimental Setup

In this study, the dataset used in this experiment is a sub-dataset from ION Corpus (Hollink et al., 2016) consisting of 60 news articles published on five newspaper websites, collected from August 2014 and August 2015. This dataset contains a collection of sentences with different difficulty level such as simple sentence which mainly the headline of a news article and the complex/long sentence which are the captions from the images.

The original news articles had to be cleaned for having a lot of information collected (title, caption, date of publication and more). The data that has been extracted from news articles primarily are headlines and captions. These text needs to go through a process of tokenization and stopword removal. The library that will be used for the NLP techniques implemented in this experiment is the Spacy Library. Both processes need to be modified as needed to produce a suitable output.

With the result from the pre-processing phase of the dataset, the texts will go through the dependency parser and a dependency tree will be produced as. Several sentences will be expressed as a dependency tree as an initial step in making rules for the rule-based relation extraction. A total of 60 articles were used to extract relationships and entities to produce triples.

To evaluate the quality and quantity of the generated triples for the multimodal KG, the evaluation is performed by manually extracting triples from the 60 articles. These sentences from the articles are examined for relevant triples and then compared to the triples extracted by the algorithm.

B. Comparison

This section will discuss the comparisons between the two methods from other works such as Gong & Wang [12] and Romadhony et al. [16]. Table V shows the summary of both the comparisons. Gong & Wang's method [12] will be compared to the way they extract only nouns to link between text as well as visuals. Romadhony et al.'s method [16] extracts the relationships that can be found from the text itself. The rules they set are from the verb and also the ADVMOD dependency relationship where it changes the predicate or verb.

TABLE V. SUMMARY OF COMPARISON

Proposed method	Gong & Wang [12]	Romadhony et al. [16]
Extract from verb and also consider preposition	only extract nouns	Extract from the verb and the ADVMOD dependency relationship
Additional attribute relationship	-	-
Has relation between image and text but also extract mostly information available	Has link between text as well as visuals but only simple text	Only extract the relationships that can be found from the text

The proposed method is an improvement of the following two works. This is because, for the method of Romadhony et al. [16], the authors focus to texts only. For [12], the method used only focuses on the image-text relationship which indirectly ignores other information that can be obtained from the text. This proposed method not only succeeds in extracting the relationship of texts, but also the relationship of image-texts also makes it very full of information to be filled in the multimodal KG to be built.

In addition, the proposed method can also produce triples of a long and complex sentence. The has-Attribute relationship makes the extracted relationship more adequate and less neglect of information from the text. This is because all the information that can be extracted from the text can be used to the fullest.

$$P = \frac{X \text{ correctly extracted triples by algorithm}}{\text{Total triples extracted by algorithm}} \quad (5)$$

$$R = \frac{X \text{ correctly extracted by algorithm}}{\text{Total relevant triples in the corpus}} \quad (6)$$

This proposed method is evaluated with the formula precision (P) and recall (R). Based on Table VI, here we can see that P = 0.90 and R = 0.60 for this study is the highest when compared to the other two methods. Thus, it can be concluded that this proposed method succeeds in extracting more triples and importantly accurate triples. The value of accuracy (P) here is high because among the 209 triples produced, there are 190 triples produced correctly. However, the recovery value (R) has only a value of 0.6 where only 190 triples are produced out of 319 triples that are theoretically capable of being produced.

TABLE VI. PRECISION AND RECALL SCORE

Total/ Method	Proposed method	(Gong & Wang, 2017)	(Romadhony et al., 2018)
Triples should be extracted (319)	The triple is extracted by the algorithm	209	71
	Triple the extracted accurately	190	60
	Precision score (P)	0.90	0.85
	Recall score (R)	0.60	0.20

This proposed method is highly dependent on parser dependency performance making it a challenge because when the performance of this parser is low as it affects the results. Generally, parser performance deteriorates with complex sentences. Another issue faced with complex sentences is the longer the sentence, the more clauses it contains, making it harder to trace back to the subject (entity) in the main clause, which resulted in the extraction of some incomplete triplets with insufficient entities, as described in the previous paragraph. Another reason for the incomplete triplets is because of co-referencing (pronouns) problem. If co-referencing resolution is performed and parser performance for complex sentences can be improved, then the value of R can also be increased.

V. CONCLUSION

The proposed multimodal KG has successfully extracted the Web image background knowledge from unstructured texts and organized in a structured graph while still maintaining the image-text relationship. A set of rules based on repeated patterns of dependency relations and POS information, can correctly extract from simple to complex sentences regardless of the domain (sport, education, world, etc.). Two additional rules are included to take into consideration the inherent correlation between the Web image with the news headline and image caption. Hence, capturing the background knowledge for Web images that are much needed by researchers in the computer vision field. This multimodal KG can be used as training data for machine learning approaches such as graph embeddings.

REFERENCES

- [1] Chan, C. S., Johar, A., & Hong, J. L. "Contextual information for image retrieval systems." Proceedings - 2013 10th International Conference on Fuzzy Systems and Knowledge Discovery, FSKD 2013, 2013, pp. 863–867.
- [2] Tiwari, P. "A Survey on Image Context Extraction Method." International Journal of Innovative Research in Advanced Engineering (IJRAE), 1(11), 2014, pp. 85–93.
- [3] Wang, B., Lin, D., Xiong, H., & Zheng, Y. F. "Joint Inference of Objects and Scenes with Efficient Learning of Text-Object-Scene Relations." IEEE Transactions on Multimedia, 18(3), 2016, pp. 507–520.
- [4] Fang, Y., Kuan, K., Lin, J., Tan, C., & Chandrasekhar, V. "Object Detection Meets Knowledge Graphs." 2017, pp. 1661–1667.
- [5] Chu, T. H., Huang, H. H., & Chen, H. H. "Image recall on image-text intertwined lifelogs." Proceedings - 2019 IEEE/WIC/ACM International Conference on Web Intelligence, WI 2019, 2019, pp. 398–402.
- [6] Vondrick, C., Oktay, D., Pirsiavash, H., & Torralba, A. "Predicting Motivations of Actions by Leveraging Text." 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 2997–3005. <https://doi.org/10.1109/CVPR.2016.327>.
- [7] Hollink, L., Bedjeti, A., van Harmelen, M., & Elliott, D. "A Corpus of Images and Text in Online News." Lrec, 2016, pp. 1377–1382.
- [8] Ramisa, A., Yan, F., Moreno-noguer, F., & Mikolajczyk, K. "BreakingNews: Article Annotation by Image and Text Processing." IEEE Trans Pattern Anal Mach Intell. 40(5), 2018, pp. 1–21.
- [9] Pan, X., Sun, K., Yu, D., Chen, J., Ji, H., Cardie, C., & Yu, D. "Improving Question Answering with External Knowledge." 2019, pp. 27–37.
- [10] Wu, Q., Wang, P., Shen, C., Dick, A., & Van Den Hengel, A. "Ask me anything: Free-form visual question answering based on knowledge from external sources." Proceedings of the IEEE Computer Society

- Conference on Computer Vision and Pattern Recognition, 2016-Decem, 2016, pp. 4622–4630.
- [11] Li, Z., Tang, J., & Mei, T. “Deep Collaborative Embedding for Social Image Understanding.” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PP(X), 2018. pp. 1. exist in the real-world is by building a knowledge graph.
- [12] Gong, D., & Wang, D. Z. “Extracting visual knowledge from the web with multimodal learning.” *IJCAI International Joint Conference on Artificial Intelligence*, 0, 2017, pp. 1718–1724.
- [13] Li, K., Zhang, J., Yao, C., & Shi, C. “Automatic relation extraction from text: A survey.” *Proceedings - 2016 International Conference on Identification, Information and Knowledge in the Internet of Things, IIKI 2016, 2018-Janua, 2016*, pp. 83–86.
- [14] Exner, P., & Nugues, P. “Entity extraction: From unstructured text to dbpedia rdf triples.” *CEUR Workshop Proceedings*, 906, 2012, pp.58–69.
- [15] Kertkeidkachorn, N., & Ichise, R. “T2KG: An end-to-end system for creating knowledge graph from unstructured text.” *AAAI Workshop - Technical Report, WS-17-01-*, 2017. pp. 743–749.
- [16] Romadhony, A., Purwarianti, A., & Widiantoro, D. H. “Rule-based Indonesian Open Information Extraction.” *ICAICTA 2018 - 5th International Conference on Advanced Informatics: Concepts Theory and Applications*, 2018, pp. 107–112.
- [17] C Carlson, A., Betteridge, J., Kisiel, B., Settles, B., Hruschka, E. R., & Mitchell, T. M.” Toward an architecture for never-ending language learning.” *Proceedings of the National Conference on Artificial Intelligence*, 3, 2010, pp.1306–1313.
- [18] Mausam, Schmitz, M., Bart, R., Soderland, S., & Etzioni, O. “Open language learning for information extraction.” *EMNLP-CoNLL 2012 - 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, Proceedings of the Conference, July, 2012, pp. 523–534.
- [19] Chaudhri, V. K. “Knowledge Graphs: What is a Knowledge Graph?.” 2021, https://web.stanford.edu/class/cs520/2020/notes/What_is_a_Knowledge_Graph.html [accessed on 10 July 2021].
- [20] Wu, P., Zhou, Q., Lei, Z., Qiu, W., & Li, X. “Template Oriented Text Summarization via Knowledge Graph.” *ICALIP 2018 - 6th International Conference on Audio, Language and Image Processing*, 2018, pp. 79–83.
- [21] Gong, D., Wang, D. Z., & Peng, Y. (2017). Multimodal learning for web information extraction. *MM 2017 - Proceedings of the 2017 ACM Multimedia Conference*, 288–296. <https://doi.org/10.1145/3123266.3123296>.
- [22] Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L., Shamma, D.A., Bernstein, M.S., & Fei-Fei, L. “Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations.” *International Journal of Computer Vision*, 2016, pp. 123, 32-73.
- [23] Farhadi, A., Endres, I., Hoiem, D., & Forsyth, D. “Describing objects by their attributes.” In *IEEE conference on computer vision and pattern recognition, CVPR 2009, 2009*, pp. 1778–1785.
- [24] Goering, C., Rodner, E., Freytag, A., & Denzler, J. “Nonparametric part transfer for fine-grained recognition.” In 2014 IEEE conference on computer vision and pattern recognition (CVPR), 2014, pp. 2489–2496.
- [25] Miao, F., Liu, H., Miao, B., & Liu, C. “Open domain news text relationship extraction based on dependency syntax.” *Proceedings of 2018 IEEE International Conference of Safety Produce Informatization, IICSPI 2018, 2019*, pp. 310–314.
- [26] She, H., Wu, B., Wang, B., & Chi, R. “Distant Supervision for Relation Extraction with Hierarchical Attention and Entity Descriptions.” *Proceedings of the International Joint Conference on Neural Networks*, 2018-July, 2018, pp. 1–8.
- [27] Cao, Q., Liang, X., Li, B., Li, G., & Lin, L. “Visual Question Reasoning on General Dependency Tree.” *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7249–7257.
- [28] Yusuf, A. A., Nwojo, N. A., & Boukar, M. M. “Basic dependency parsing in natural language inference.” 2017 13th International Conference on Electronics, Computer and Computation, ICECCO 2017, 2018-Janua, pp. 1–4.
- [29] De Marneffe, M. C., Dozat, T., Silveira, N., Haverinen, K., Ginter, F., Nivre, J., & Manning, C. D. “Universal stanford dependencies: A cross-linguistic typology.” *Proceedings of the 9th International Conference on Language Resources and Evaluation, LREC 2014, 2014*, pp. 4585–4592.
- [30] Choi, J. D. “ClearNLP Dependency Labels.” https://github.com/clir/clearnlp-guidelines/blob/master/md/specifications/dependency_labels.md, 2015, [accessed on 20 October 2021].
- [31] A Altinok, D. “Mastering spaCy (1st ed.)” Packt Publishing. <https://www.perlego.com/book/2742862/mastering-spacy-pdf>, 2021, [accessed on 20 October 2021].

Soil Color as a Measurement for Estimation of Fertility using Deep Learning Techniques

N Lakshmi Kalyani, Kolla Bhanu Prakash

Department of CSE, Koneru Lakshmaiah Education Foundation
Guntur (DST), Andhra Pradesh, India

Abstract—Soil Behavior helps the farmer predict performance for growing crops, nutrient movement, and determine soil limitations. The traditional methods for soil classification in the laboratory require time and human resources and are expensive. This analysis examines the possibility of image recognition by artificial intelligence, with a machine learning technique called deep learning, to develop the cases that use artificial intelligence. This study performed deep learning with a model using a neural network. Neural Networks has used to evaluate relationships between the parameters of the three-dimensional coordinates resulting in soil classification and parameters. So Artificial Neural Networks (ANN) can be an effective tool for soil classification. This paper focused on AI techniques used to predict the soil type, advice the crop to yield, and discuss the transformed learning and benefits.

Keywords—Artificial neural networks; deep learning; soil classification; soil nutrients; data augmentation; transform learning

I. INTRODUCTION

Agriculture stands very quintessential in society. Agriculture is a source of livelihood in most parts of the world. Agricultural produce is of great importance. But in recent years, the farm produce is gradually decreasing. Soil plays a crucial role in agriculture. Soil consists of nutrients that the plants use to grow. There exist various kinds of soils available, and each has different effects. Crop's productivity was mainly based on the type of soil [1]. The possible way to improve productivity is that choose the right crop for the right land type. Soil needs to be investigated primarily before classifying them into distinctive groups. Based on these soil groups and the geographical conditions, one can decide which crop is best suited and is beneficial. The traditional methods are Costly, lengthy process and also time-consuming.

Consequently, there is a necessity for new technologies and methods to enhance the existing system to get faster and better results. Soil texture is due to the percentage of silt, clay, and sand present in it, and change in the fraction may result in different colors due to interaction among soil characteristics. Classification has a significant influence in agriculture for estimating yielding crops, and soil classification will define the relation between soil samples and other natural substances present. Convolutional Neural Network (CNN) is the branch of Deep learning and enables the machine to exhibit self-learning, to show the intelligence to predict by analyzing the input [3]. The essential point here is the privilege of the computer to learn automatically without human interaction.

Soil is the source of the minerals, liquids organisms that produce the foundation for the plant, and classification plays a major role in managing a crop, increasing soil primarily before classifying them into distinctive groups. Based on these soil groups and the geographical conditions, one can decide which crop is best suited and is beneficial [4]. The traditional methods are Costly, lengthy process and also time-consuming.

This research aims to benefit the former in identifying the red soil, and it analyzes the different soil patterns using CNN and recommends particular crops in that soil. And also help the researcher in soil science. The general purpose of this research is to compare the different soils and identify the red soil. The consequence of this may have numerous advantages to agriculture, soil management, and the environment.

II. LITERATURE SURVEY

In this section, we summarized various works related to machine learning, image processing, and classification models from different papers seen in Table I.

The least median squares regression techniques produce more reliable results than the classical linear regression technique from the set of characteristics [2].

SVM classifier can work efficiently with high level of accuracy. MATLAB software has proved an efficient tool for design and development of classifier and can be used for further development of independent interface for on-site real time soil classification [5].

Time-honored methods of soil assortments are standard penetration test (SPT), cone penetration test (CPT), pressure meter test (PMT), and vane shear test (VST), time-consuming and needed for accurate results. Soil classification can be automated using artificial neural network techniques [6].

Mrs.Saranaya and Ms.A.Mythili researched to analyze the soil types so that it is helpful to farmers to choose a crop that to be cultivated. They have considered the SVM algorithm and chemical effects of soil like pH, salinity, organic matter, potassium, sulphur, zinc, Boron, calcium, Magnesium, Copper, Iron, and Manganese [7].

Classification of soil and quality prediction using a software-enabled solution using machine learning algorithms including decision tree, by considering soil chemical parameters [8].

TABLE I. VARIOUS MODELS USED FOR SOIL CLASSIFICATION

S.No.	Title	Techniques
1	Soil classification and crop suggestion using machine learning	Naive Bayes, J48, JRip algorithms
2	An Intelligent Model for Indian Soil Classification using various Machine Learning Techniques	The Pre-processed images are feature extracted, and the data extracted is used to train the SVM classifier.
3	Recent Trends Of Machine Learning In Soil Classification	Various emerging machine learning algorithms like SVM's, KNN, ANN, DT(decision tress) were discussed.
4	Soil Classification and Crop Suggestion using Machine Learning	Bagging Classifier, SVMachine, and KNN for classification of soil and crop recommendation.
5	Machine Learning in Soil Classification and Crop Detection	Image acquisition, Pre-processing, KNN, Feature extraction, SVM classifier.
6	Prediction Of Soil Quality Using Machine Leaning Techniques	Decision Tree and Random forest algorithms
7	Artificial intelligence system for supporting soil classification	Neural network models are used, CNN, steepest descent method.
8	Soil Classification & Characterization Using Image Processing	Technically proposed system has been based on HSV, Enhancement algorithm, and SVM classification algorithm.
9	Soil Physical Properties	Stokes regulation, bulk density, particle density, mass flow
10	Chemical properties of soils	Different types of chemical organic molecules.
11	Nutrients Detection in the Soil: Review Paper	Soil Nutrient detection, Reflectance sensing, Electro chemical sensing, Electro Conductivity sensing.
12	Soil Classification Methodology: Critical Analysis	Classification Methods, Inventory Methods.
13	Soil Data Analysis Using Classification Techniques and Soil Attribute Prediction	Data mining, classification, regression, soil testing, agriculture , WEKA Tool.
14	Soil Classification & Characterization Using Image Processing	Classification of soil by image processing using SVM technique, Machine Learning.
15	Soil Classification using Machine Learning Methods and Crop Suggestion Based on Soil Series	Classification of soil by image processing using SVM technique based on colour, energy, HSV of soil.
16	Determination of Soil Nutrients and pH level using Image Processing and Artificial Neural Network	Rapid soil Testing, Soil Test kit, Artificial Neural Networks, Image processing are used.
17	Soil Quality Measurement using Image Processing and Internet of Things	IOT and image processing to measure pH, moisture present, Soil Nutrients. Raspberry-Pi and camera are used.
18	Analysis of Soil Behaviour and Prediction of Crop Yield using Data Mining Approach	Naive Bayes and K- Nearest Neighbour methods are used.

Prediction soil type was performed by deep learning model using neural network. Research carried out for three classes of soil, specifically for clay, sand, and gravel, an AI model, was constructed that was keen efficient homogeneity of the images utilized and the research has shown that this AI model can be applied to conceive judgments on soil classification [9].

The SVM classifier was found to outperform over all other techniques. Maximum Likelihood classification, sub-pixel classification, and ANN classification were chosen as image classification techniques from various methods [11].

Soil classification and segmentations are necessitated, along with soil nutrients to recommend crops and fertilizers to use. Non-specialists farmers can understand features easily [10].

Researcher T.Abimala proposed a model to predict the variety of soil applying image processing techniques. Classification of soil is by the processing of soil image color and texture patterns [12].

Soil Characteristics have a notable impact on the response of soil in the field of agricultural uses. Soil aeration depends on the soil texture, is a static property of soil quality. Plants' growth depends on soil quality and temperature [13]. Weathering processes may create porous media at the land surface. The correlation between soil and bio-indication may not be strong enough to bring out such suitable changes. Soluble elements are purged into the lower layers of soils where they accumulate. Insoluble chemical elements remain in the upper layers of the soil [14].

Soil nutrients and pH define through image processing and ANN [15]. The process of soil classification systems based solely on grain diameter is capable of misunderstanding because the physical properties of outstanding soil elements depend on many factors other than grain size [16]. There is also the quantity of water (humidity) which is also a non-negligible factor. It is the newest and most used because it is the improved form of all others [17].

The least median squares regression can predict soil is identified to generate abler results than the classical linear regression by analyzing soil sample and cropping pattern [18].

Image processing using SVM techniques classify the soil based on color, energy, and HSV [19]. The researcher recommended a suitable crop for a particular soil by analyzing the PH value, moisture level, and nutrients collected using sensors with IoT technology. Here former need to have a piece of knowledge on hardware [20].

Experimental study says yield prediction is possible with AI Techniques. ANN attained more accuracy than RF, MLR, and NB Techniques [21]. Categorization of soil could be done using Data mining techniques like K-Nearest neighbors and naive Bayes as low, medium, and high. It helps to choose suitable land for more profitable crop production [22]. SVM, principal component analysis can be added further.

The CNN-based regression model is developed based on the characteristics of the TIR images, with required generality and performance. This model, in comparison with DNN, gives

promising accuracy. Remote digital imaging using a drone is advantageous than data acquired using satellite [23].

There are non-exhaustive methods utilizing image processing, machine learning, and deep learning in previous research work related to soil classification [24]. And many researchers focused on traditional methods, image processing, computer vision techniques, and computer-based applications. This paper proposed advanced soil classification techniques using machine learning and deep learning. The primary purpose of soil category is to forerun behavior, determine the uses, assess their productivity, and extrapolate analysis to predict nutrients.

This paper discusses the architecture of the proposed system, methodology, results, concluding remarks, and future scope.

III. METHODOLOGY

In Telangana, 48% of the land covered red soil and types of red soil and color and causes listed in Table II.

Some of these soils look reddish due to the wide diffusion of iron in crystalline and metamorphic rocks, and some are yellow due to diffusion occurring in a hydrated form. Soil is the foothold for plants' roots and holds the necessary nutrients for plants to grow. It filters rainwater and regulates the discharge of excess rainwater, preventing flooding, and it is capable of storing large amounts of organic carbon. All these depend on the characteristics of the soil. The primary skill to succeeding in farming here understands soil type and giving the most fitting plants for that soil [14].

Soil classification could be made as per engineering properties, and it could be free from trouble-free from field survey and the usually, engineers classify soils as per soil characteristics. Classification would commence from the soil image. Convolution Neural Network is an incredible image recognition invention and is a sub-class of deep learning neural networks. CNN generally used to analyze images and image classification.

CNN Architecture Fig. 1 has four layers listed as follow:

- Convolution layers.
- ReLu layers.
- Pooling layers.
- Fully connected layers.

TABLE II. RED SOIL CLASSIFICATION

S.No	Soil Type	Colour	Chemical
1	Red Clay Soil	Red	Iron oxide
2	Red loam soil	Red	Potash
3	Red Laterite Soil	Red	Iron and Aluminium
4	Red –Yellow soil	Red-Yellow	Ferric hydroxide
5	Red Sandy soils	Red	Iron
6	Red Gravel Soil	Red	Iron

A. Convolution Layer

- Convolutional layers: these layers apply a convolution operation to the input image and pass the information on to the next layer.
- Pooling Layer: The next layer is the pooling layer, where outputs combined as cluster of neurons into a single layer. And the next layers are fully connected layers, in which all neurons connect to each neuron in the next layer.

B. Relu

Proposed CNN Model and trained it on a Soil image dataset. The results observed that performance was minimal. Training a model with millions of images takes days to achieve high performance in real-world applications. An alternative is to use a pre-trained model, and it would retain on our dataset as transform learning.

In this Model, an image was directly given to the algorithm, which will classify the given image as either red soil or not, as shown in Fig. 1 and the detailed flow explained in Fig. 2.

C. Pre-Processing

A smoothing (low pass) filter has been used to eliminate high-frequency noise and artifacts from the picture. Smoothing filters employ a moving window operator that adjusts the value of one pixel of an image at a time based on a function of a local area of pixels. As the operator advances over the picture, all of the pixels are affected. As a consequence, the smoothing filter progressively enhances the image over time by eliminating imperfections.

D. Feature Extraction

The feature extraction stage is the essential phase in the process. It encompasses all of the characteristics that are needed to classify the soil type, such as texture, color, and intensity was extracted. As a result, a metric known as color moments was employed to distinguish photographs based on their color characteristics.

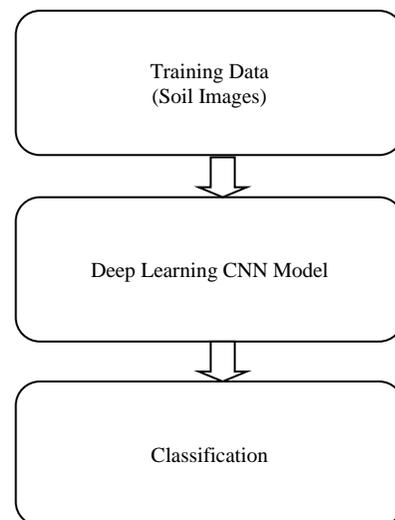


Fig. 1. Abstract Level Soil Classification Architecture.

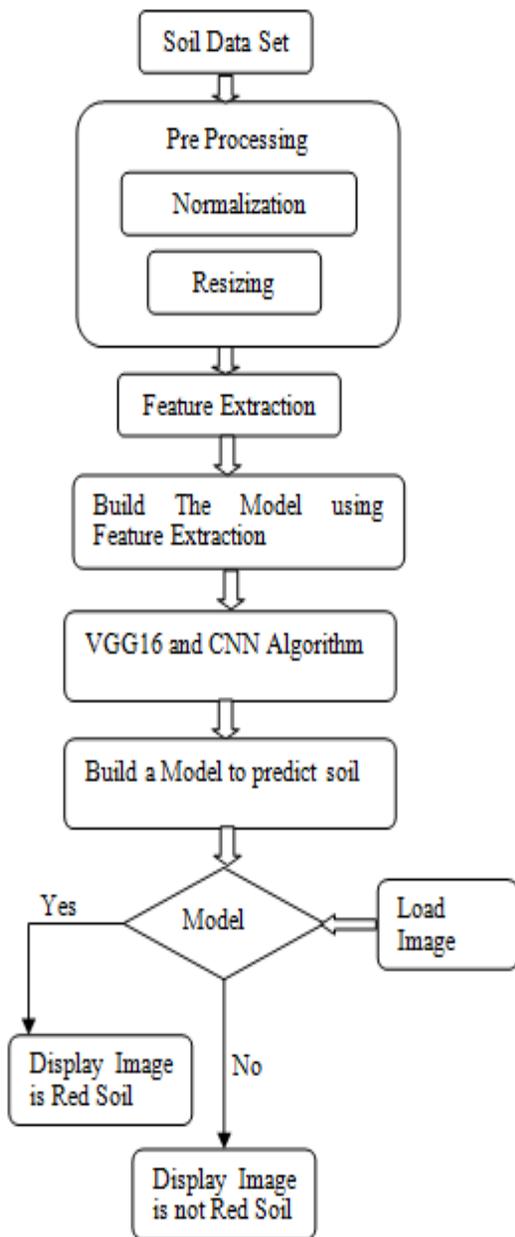


Fig. 2. Detailed Soil Classification Architecture.

E. CNN Classification Model

CNN or ConvNet are deep neural networks used for image recognition, object detection, and classification. Image classification is the process of deciding which class (or combination of categories) best describes an input image. In CNN, we take a picture as an input, assign weight to the image's various aspects/features, and differentiate one from the other. CNN requires much less pre-processing than other classification techniques.

F. Softmax

Softmax is an intriguing activation function since it maps our output to the [0, 1] range and maps each output so that the entire sum equals one. Softmax's output is a probability distribution as a result as shown in Fig. 3.

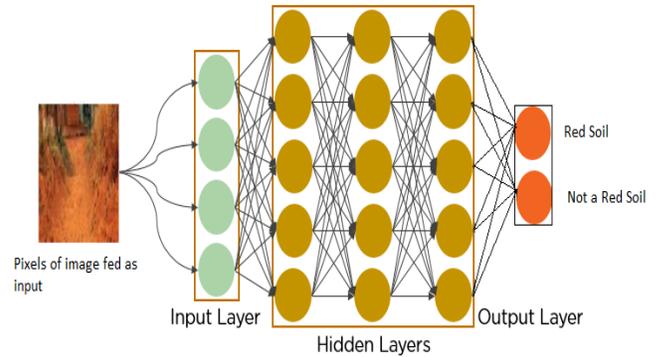


Fig. 3. Soil Classification with CNN.

In-Training the Model, all the images are converted into arrays and stored in Data Variable. Each image is given labels for the detection and stored in Variable Labels.

After preprocessing the steps to be done to build the model Average pooling is applied to down sample the input dimension, Height, and width by taking average value over a window of size 4x4 for each input channel. Strides shift the window along each dimension.

Flatten transforms the pooled feature pictures into a single column, then sent to the fully linked layer. Dense added the fully connected layer to the convolution neural network, and dropout is a strategy for preventing over fitting in a model. Once the model has fitted with the Layers, the trainable parameters have been set.

Fig. 4 shows how many total parameters are there and how many are available for the training.

```

In [19]: bmodel.summary()

Model: "vgg16"
-----
Layer (type)                Output Shape              Param #
-----
input_1 (InputLayer)        [(None, 224, 224, 3)]    0
block1_conv1 (Conv2D)       (None, 224, 224, 64)     1792
block1_conv2 (Conv2D)       (None, 224, 224, 64)     36928
block1_pool1 (MaxPooling2D) (None, 112, 112, 64)     0
block2_conv1 (Conv2D)       (None, 112, 112, 128)    73856
block2_conv2 (Conv2D)       (None, 112, 112, 128)    147584
block2_pool1 (MaxPooling2D) (None, 56, 56, 128)      0
block3_conv1 (Conv2D)       (None, 56, 56, 256)      295168
block3_conv2 (Conv2D)       (None, 56, 56, 256)      590080
block3_conv3 (Conv2D)       (None, 56, 56, 256)      590080
block3_pool1 (MaxPooling2D) (None, 28, 28, 256)      0
block4_conv1 (Conv2D)       (None, 28, 28, 512)      1180160
block4_conv2 (Conv2D)       (None, 28, 28, 512)      2359808
block4_conv3 (Conv2D)       (None, 28, 28, 512)      2359808
block4_pool1 (MaxPooling2D) (None, 14, 14, 512)      0
block5_conv1 (Conv2D)       (None, 14, 14, 512)      2359808
block5_conv2 (Conv2D)       (None, 14, 14, 512)      2359808
block5_conv3 (Conv2D)       (None, 14, 14, 512)      2359808
block5_pool1 (MaxPooling2D) (None, 7, 7, 512)        0
-----
Total params: 14,714,688
Trainable params: 14,714,688
Non-trainable params: 0
  
```

Fig. 4. Listing of Parameters in Training.

IV. RESULT

Every step picks images from the training set at random. Then, each step finds its bottlenecks from the collection. And ultimately, each step puts the images into the final layer to get predictions. Then, the forecasts are equated alongside the actual labels to inform the final layer's weights using the back-propagation procedure.

Each step shows the level of training accuracy and validation accuracy Fig. 5. Training accuracy was the prediction of training images that were classified correctly. Quality of the model measured by validation accuracy and to what extent it guess depending on data it has not seen before. Suppose 86% percentage on the training set and 85% on the validation set. You can expect your model to perform with 85% accuracy on new data.

From Fig. 6 Prediction Accuracy into a percentile, the accuracy of the prediction soil image classifier is 91%.

This model could test using User Interface (UI). Once the model was ready, we developed a web application using the Flask framework. Here users can upload images to predict soil type.

```

Compiling Starts
Epoch 1/10
39/39 [=====] - 89s 2s/step - loss: 0.1968 - accuracy: 0.5649 - val_loss: 0.5901 - val_accuracy: 0.812
5
Epoch 2/10
39/39 [=====] - 92s 2s/step - loss: 0.1955 - accuracy: 0.7078 - val_loss: 0.5456 - val_accuracy: 0.737
5
Epoch 3/10
39/39 [=====] - 92s 2s/step - loss: 0.1172 - accuracy: 0.7922 - val_loss: 0.4793 - val_accuracy: 0.812
5
Epoch 4/10
39/39 [=====] - 92s 2s/step - loss: 0.1848 - accuracy: 0.8019 - val_loss: 0.4508 - val_accuracy: 0.825
0
    
```

Fig. 5. Training the Model.

	precision	recall	f1-score	support
Soil_Dataset//Black_Soil	0.93	0.91	0.92	43
Soil_Dataset//Red_Soil	0.89	0.92	0.91	37
accuracy			0.91	80
macro avg	0.91	0.91	0.91	80
weighted avg	0.91	0.91	0.91	80

Fig. 6. The Accuracy of the Model.

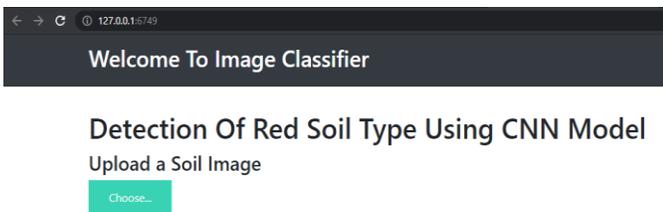


Fig. 7. UI of Web Application.

Fig. 7 demonstrates the web application with one button to upload an image of soil. After uploading an image predicted labeled button will appear on the screen, as shown in Fig. 8. User needs to click on predict button and then display the result as RED Soil as shown in Fig. 9. The exact process is repeated with a different image and displays the result as NOT a RED SOIL, as shown in Fig. 10.



Fig. 8. Uploading Soil Image.



Fig. 9. Prediction of Soil Type.



Fig. 10. Testing Application with other Soil Image.

V. CONCLUSION

The proposed model categorizes soil using CNN successfully identified the Red soil and tested in the field got 91 percent accurate results. And can estimate the soil's fertility by predicting significant nutrients for specific soil. Introducing cutting-edge technologies in Agriculture can improve the yield by applying adequate application fertilizers, and with an improved dataset, also get 99 percent accurate predictions.

REFERENCES

[1] N Lakshmi Kalyani, Kolla Bhanu Prakash, "Soil Synthesis and Identification of Nitrogen percentage in Soil using Machine learning algorithms and Augmented Reality – A Typical review", IJETER, 2020(8), PP.5501-5505.

- [2] Shravani.V,S.Uday KiranG, "Soil classification and crop suggestion using machine learning," IRJET,2020.
- [3] Prakash K.B., Dorai Rangaswamy M.A. Content extraction studies using neural network and attribute generation, 2016, Indian Journal of Science and Technology, 9-22, 1(10).
- [4] Prakash K.B. "Information extraction in current Indian web documents",2018,International Journal of Engineering and Technology(UAE),7,2.8,68-71.
- [5] Chandan , Ritula Thakur', "An Intelligent Model for Indian Soil Classification using various Machine Learning Techniques," ICER, 2018.
- [6] Dido, A.A., Krishna, M.S.R., Singh, B.J.K., Tesfaye, K., Degefu, D.T. "Assessment of variability of yield affecting metric characters in barley (*Hordeum vulgare*) landraces " Research on Crops, 2020, 21(3), pp. 587–594.
- [7] Mrs. N.Saranya,Ms. A. Mythili, "Classification of Soil and Crop Suggestion using Machine Learning Techniques," IJERT,2020.
- [8] Ashwini Rao ,Janhavi U ,Abhishek Gowda N ,S Manjunatha, Mrs.Rafega Beham "Machine Learning in Soil Classification and Crop Detection," IJSRD, 2018.
- [9] Shinya Inazumi, Ph.D. , Sutasinee Intui, M.Eng. , Apiniti Jotisankasa, Ph.D. , Susit Chaiprakaikeow, Ph.D. , Kazuhiko Kojima, "Artificial intelligence system for supporting soil classification," ELSEVIER, 2018.
- [10] Hement Kumar Sharma, Shiv Kumar, "Soil Classification & Characterization Using Image Processing," IEEE , 2018.
- [11] Priyanka Dewangan, Vaibhav Dedhe, "Soil Classification Using Image Processing and Modified SVM Classifier", IJTSRD, 2018.
- [12] Vijaya Nagini, D., Krishna, M.S.R., Karthikeyan, S. "Identification of novel dipeptidyl peptidase-iv inhibitors from ferula asafoetida through gc-ms and molecular docking studies," IJPT, 2020, pp.5072-5076.
- [13] Sk Al Zaminur Rahman, Kaushik Chandra Mitra, S.M. Mohidul Islam, "Soil Classification using Machine Learning Methods and Crop Suggestion Based on Soil Series," IEEE, 2018.
- [14] A. Balasubramanian,CHEMICAL PROPERTIES OF SOILS,2017.
- [15] John Carlo PUNO , Edwin SYBINGCO , Elmer DADIOS , Ira VALENZUELA , Joel CUELLO, "Determination of Soil Nutrients and pH level using Image Processing and Artificial Neural Network," 2017.
- [16] Ruwali A., Kumar A.J.S., Prakash K.B., Sivavaraprasad G., Ratnam D.V. Implementation of Hybrid Deep Learning Model (LSTM-CNN) for Ionospheric TEC Forecasting Using GPS Data,2021, IEEE Geoscience and Remote Sensing Letters,18(6),9093827,pp:1004-1008.
- [17] Mbuya Mukombo Jr. , Mutonkole Ngomba H., Musimba Kasiya A , Ngoy Biyukaleza B., "Soil Classification Methodology: Critical Analysis," IJSR, 2018.
- [18] Kunal Teeda, Nandini Vallabhaneni, Dr.T.Sridevi Jay Gholap, "Comparative Analysis of Data Mining Models for Crop Yield by Using Rainfall and Soil Attributes," IEEE, 2018.
- [19] Hement Kumar Sharma, "Soil Classification & Characterization Using Image Processing," IEEE,2018.
- [20] Dileep Reddy Bolla, Dr.Shivashankar, Anirudh Sandur, "Soil Quality Measurement using Image Processing and Internet of Things," 2019, IEEE.
- [21] Vaishali Pandith , Haneet Kour , Surjeet Singh , Jatinder Manhas , and Vinod Sharma, "Performance Evaluation of Machine Learning Techniques for Mustard Crop Yield Prediction from Soil Analysis," JSR, 2020.
- [22] Monali Paul, Santosh K. Vishwakarma, Ashok Verma, "Analysis of Soil Behaviour and Prediction of Crop Yield using Data Mining Approach," IEEE, 2015.
- [23] Bharadwaj, Prakash K.B., Kanagachidambaresan G.R. "Pattern Recognition and Machine Learning", 2021, EAI/Springer Innovations in Communication and Computing, 105-144.
- [24] Prakash K.B. "Content extraction studies using total distance algorithm", 2017, Proceedings of the 2016 2nd International Conference on Applied and Theoretical Computing and Communication Technology, iCATccT 2016, 7912085, 673-679.

A Survey on Genomic Dataset for Predicting the DNA Abnormalities Using ML

Siripuri Divya¹, Y. Bhavani², Thota Mahesh Kumar³

M.Tech, Data Science, Kakatiya Institute of Technology & Science Warangal, India¹

Associate Professor, Dept. of IT, Kakatiya Institute of Technology & Science, Warangal, India²

Assistant Professor, Dept. of IT, Kakatiya Institute of Technology & Science, Warangal, India³

Abstract—Genomic data is used in bioinformatics for collecting, storing and processing the genomes of living things. In order to process the genetic information, machine learning algorithms plays a vital role in building a computational model by using the statistical theory. This paper helps the researchers, who are doing research with the DNA dataset by applying the machine learning logics. Feature scaling machine learning techniques helps in predicting the sequence of genome for extrachromosomal amplification and predicting the tumor intensity in the human gene. Identification of unconventional chromosome in the DNA sequence minimizes the structural risk. In this paper, researchers can get clear insight on classification, sequence prediction, fuzzy relationship and SNP on genome dataset. The performance of various existing models is measured using the performance metrics and the accuracy.

Keywords—Genomic data; deoxyribonucleic acid (DNA); machine learning algorithms; single nucleotide polymorphism (SNPs)

I. INTRODUCTION

Deoxyribonucleic acid (DNA) is the combination of different nitrogenous bases, phosphate molecule and sugar molecules which are combined to form nucleic acid. These nucleic acid molecules consist of genetic information that are used for transmitting organic material from parent to child. All the combinations in the nucleic acid of a cell are arranged in a sequence to form a DNA structure. DNA are mainly responsible for storing genetic information and also for producing the proteins in human body through transcription and translation process. DNA are in double helix structure, present in eukaryotic and prokaryotic cells.

Deoxyribonucleic acid (DNA) is located in each and every cell nucleus of human body containing the genetic information. The cell nucleobases of DNA consist of four nitrogen nucleotides namely adenine represented as A, cytosine represented as C, guanine represented as G and thymine represented as T formed using the nitrogen bond as represented in the Fig. 1. DNA of a cell nucleus consist of chromosomes. There are 3-billion pair of chromosome sets termed as genome. Genome consist of 46 chromosomes in the DNA sequence grouped to 23 pairs.

DNA can be divided into four different categories based on the structure as A-form, B-form, C-form and Z-form DNA. The base pairs which are not in perpendicular to the helix axis are defined as the A-form DNA, these protect the human body in the extreme desiccation effect of bacteria. The

base pairs which are perpendicular to the helix axis form are defined as the B-form DNA, these are responsible for gene expression, mutation under normal conditions. The zig-zag form of base pairs to the helix axis are defined as Z-form DNA, these are responsible for gene regulation. The base pairs which are in the form of non-integral helix structure are defined as the C-form DNA, these are responsible for cloning the eukaryotic genes.

DNA consisting of 3 billion base pairs carrying information from one gene to another gene are created and replicated based on the human genetics and the life-style. The DNA arranges the base pairs in a chain sequence to perform transcription and translation process to produce protein in mRNA to develop the human body. These proteins are used as energy to do large amount of work by our body. The DNA structure is divided into two halves or two single strands. The strands in DNA are thin molecules wrapped to form a helix structure, these are composed of blocks or nucleotides. Human genome is around 98% similar to chimpanzees and 75% similar to mouse.



Fig. 1. Deoxyribonucleic Acid [23].

A. DNA-Sequence-Classification

DNA can be classified based on various standards like structure, number of base pairs, location, coiling patterns, nucleotide sequences, number of strands, coding and non-coding. DNA sequence-classification is performed based on the nucleotide combinations in the sequence. The four nucleotides A, T, C and G forms a series in the nucleic acid of the cell. The process of identifying the nucleotide sequence in the DNA are used to determine the order of nucleic acid sequence. DNA sequences are classified to get the information related to the evolution of various species, living organisms and their transformations, medical-diagnoses, forensic

investigations and organism identification. The nucleotide canonical DNA structure using the computer terminology can be classified for various predictions like disease risk, next generation sequence, cancer research, birth defect screening, food importing/exporting control, paternity testing, drug target and gene therapy.

The process in which the information from the DNA is carried out to the RNA molecules are called as transcription. Transcription process is carried out in three different stages in the gene expression, enzymes that perform transcription in copying the DNA strand from the DNA [24] sequence in eukaryotes cell is called RNA polymerases. The single stranded DNA correlation with the complementary strand RNA is performed using the RNA polymerases by adding the new-nucleotides. The first step in the gene transcription is initiation of promoters for binding the RNA polymerase with the DNA sequence molecules in each gene. The second step is elongation process where the RNA molecule builds the complementary nucleotides chain in the template strand. Termination is the final step in the transcription where the sequence mechanism is formed as hairpin RNA molecule.

The process in which the information from the termination of RNA molecule is translated to the ribosome to produce proteins is called translation process. In the translation process the genetic code from the RNA molecule is converted to the amino acid 20-letter code to produce the protein blocks.

The translation in the ribosome is performed in three different stages. The first step of gene translation is initiation, where the small-ribosomal-subunit binds the information from the transcription and codons initializes the methionine code & AUG to transfer the information. The second step of translation is elongation where the codon continues to increase the chain by adding the corresponding amino-acid using the peptide bond. The final step in the translation process is the termination where the proteins are produced by completely binding the codons from the RNA molecule.

Fig. 2 displays the transcription control flow and the translation process, in converting the DNA information to produce the proteins. Transcription synthesis of single stranded RNA from a double stranded DNA template is used to produce messenger RNA. Translation is the first stage of protein biosynthesis from RNA in the gene expression.

B. DNA Methylation

DNA Methylation is the process in which the methylation activity in the DNA segment is changed without changing the original gene sequence. Methylation is a process where the methyl groups are attached to the DNA molecule to repress-gene. DNA methylation occurs during the epigenetic event. The covalent modification of DNA methylation results in three types of methylated bases called C5-methylcytosine(5mC), N4-methylcytosine(4mC), N6-methyladenine(6mA). The DNA methylation is important for transcriptional-gene-silencing, genomic imprinting, maintaining the genome stability, embryonic development and X-chromosome inactivation's.

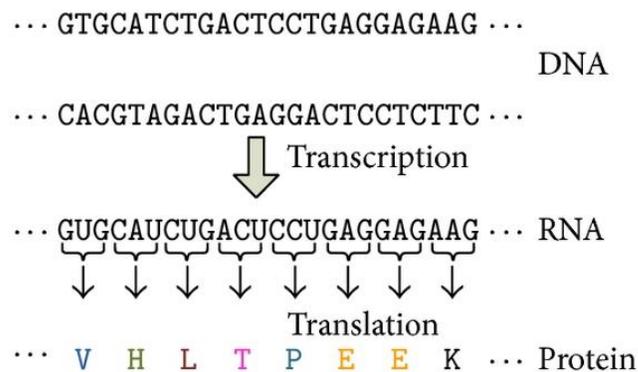


Fig. 2. DNA Transcription and Translation.

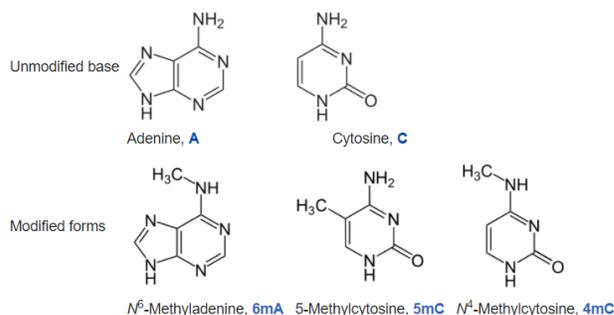


Fig. 3. DNA Methylation.

Fig. 3 is the chemical modification of the cytosine, where by addition of a methyl group to the number 5 carbon of the cytosine is converted to 5-methylcytosine which is followed by the guanine dinucleotide CpGs process.

C. DNA Damage

DNA damage is the process of alteration in the DNA structure resulting in the chemical abnormalities. DNA damage are mainly caused due to the change in the environmental factors and the metabolic process inside the cell. The main source of the DNA damage is endogenous damage with in the cell and exogenous damage caused by the external agents like X-rays, UV-rays. The DNA damage can be classified into three different types based on the alteration in the genetic material as single-base-alteration, two-base-alteration and chain-breaks-and-cross-linkages. The abnormality in single base of the DNA is caused by depurination, alkylation, deamination and base-analog formation. The two base alteration is caused by UV induced dimer formation in the thymine and bifunctional-alkylating agent. The chain-breaks-and-cross-linkages are caused by ionizing radiation, oxidative-free-radical-formation, radioactive disintegration, cross linking-between bases in same or opposite strand, cross linking between DNA and protein molecules. The Fig. 4 represents the DNA damage caused during the cellular alteration. Agents damaging the DNA [22] are radiations caused due to highly reactive oxygen radicals, ultraviolet rays and ionizing radiations, chemicals in the environment like aromatic hydrocarbons and aflatoxins.

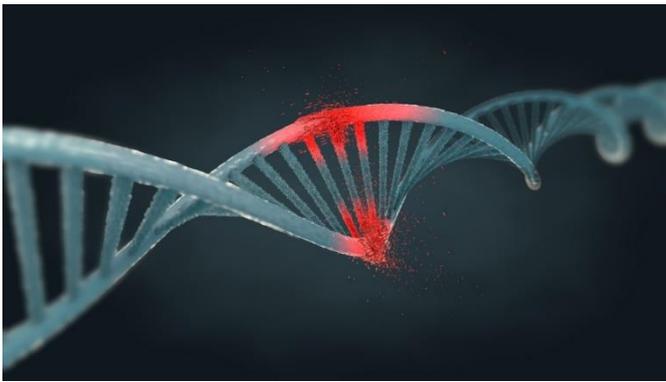


Fig. 4. DNA Damage [26].

DNA repair can be performed by the cell in two ways, direct-damage-reversal and excision of DNA damage. In direct-damage-reversal initially the single polypeptide with the enzymatic properties binds the chain and restores using DNA photolyases and alkyl transferases. In the excision of DNA damage, it solves the excised free bases generated by altering the bases to deoxyribose-phosphate by initializing the DNA glycosylases called base-excision-repair. Nucleotide-excision-repair mechanism is used to replace the DNA damage in 30 bases. Mismatch repair mechanism allows the enzymes to identify the strand and replace them with normal cellular enzymes corresponding to the base-pair-rules, strand break repairs, single-strand breaks and double-strand break damages. The diseases caused due to the defect in the DNA repair system are ataxia telangiectasia, bloom syndrome, Cockayne's syndrome, progeria syndrome, rothmund-thomson syndrome, trichothiodystrophy, Werner syndrome, xeroderma pigmentosum and hereditary non polyposis colon cancer.

Fig. 5 represents the Nucleotide-excision-repair mechanism is used to replace the DNA damage in 30 bases.

D. Mitochondrial DNA

Mitochondrial DNA is responsible for the cellular-metabolism, oxidative-stress-control and apoptosis. Mitochondria is also known as the power-houses of DNA cell, it is inherited from the mother's ovum. It consists of 13 coding genes and the 24 non-coding gene of length 16,569 bp in the human body. Mitochondria uses the oxygen and sugars to create energy of the main cell. The mitochondria mainly designed with 22tRNA and rRNA coding genes to control replication and transcription process in the cell. It is in circular structure with several copies of single mtDNA-molecules present freely in the nuclear envelope.

Fig.6 Mitochondrial organelles structure are found in the cell cytoplasm as represented in Fig. 6 which consist of some of components like 2 membranes called inner membrane and outer membrane protecting the cellular matrix.

E. DNA Mutation

The heritable change in the arrangements of genetic material chromosomes position is termed as mutation or DNA mutation. Mutations are occurred in gametes and causes permanent change in the genetic sequence of nucleotide

forming new amino acid. The gene mutation rate is 1 or 2 new mutations in 1000000 genes during the DNA copy. Mutations are unpredictable and can be of many forms like gene mutation or point mutation and chromosome-mutation. These mutations are caused due to change in the complete chromosomes structure, chromosomes count and single pair of chromosome's structure. Some of the forms of mutations are due to the addition of extra nucleotide that causes gene mutation and addition of extra chromosome that causes chromosome mutation. Deletion of nucleotide chain from the gene sequence causes gene mutation and chromosomes are lost in the gene sequence causing chromosome mutation. Duplication of nucleotide chain is repeated in gene mutation and chromosomes are repeated in chromosome mutation. In inversion nucleotide sequence are detached from the gene sequence causing gene mutation and deleted chromosomes rejoining the chain in the inverse position causing chromosomes mutation.

Fig. 7 represents different mutations caused in the DNA. The first sequence in the Fig. 7 represents the normal gene sequence consisting of cytosine, thymine, adenine and guanine. The insertion of new guanine in the sequence at the second position results in the mutation. The deletion of adenine in the 3rd position of the sequence, duplication of cytosine and thymine, inversion of the 2nd and 3rd position of the sequence causes DNA mutation.

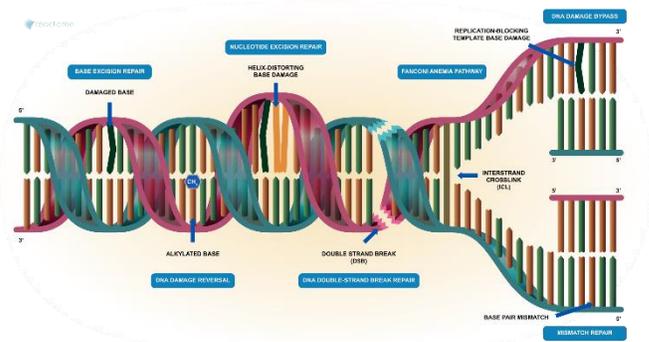


Fig. 5. DNA Repair [25].

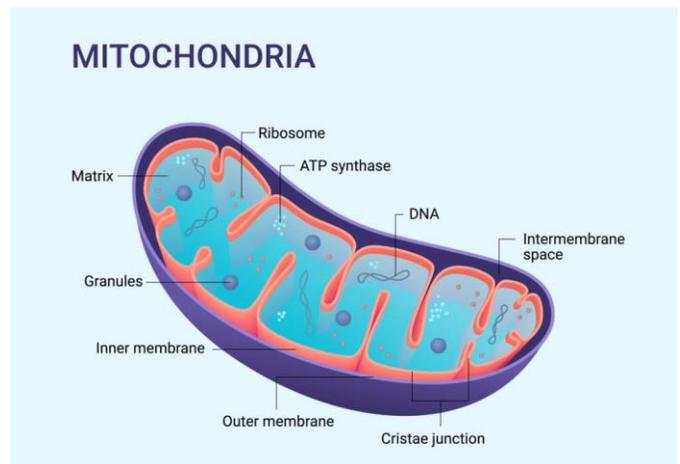


Fig. 6. Mitochondria.

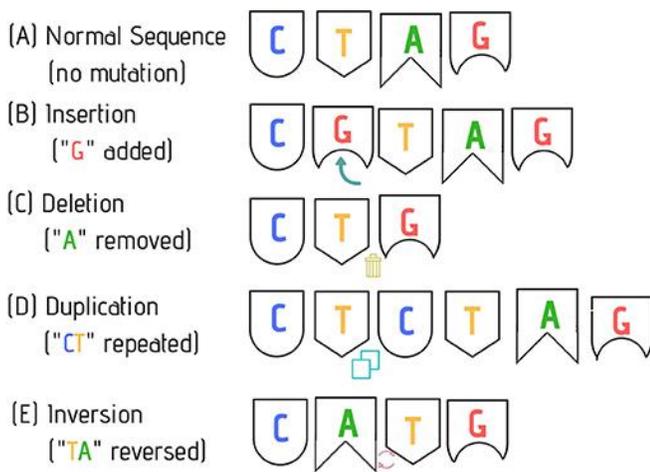


Fig. 7. DNA Mutation.

II. BACKGROUND AND MOTIVATION

Deoxyribonucleic-acid composed of 46 chromosome's carries genetic information from parents to children in homo sapiens. Chromosomes are made up of different groups of purine-pyrimidine bases, phosphate and sugar attached to the one carbon of deoxyribose's forming as adenine represented as A, cytosine represented as C, guanine represented as G and thymine represented as the T formed using the nitrogen bond. Human genome is used for scientific research for understanding the physical gene sequence and base pairs functioning. DNA research for analyzing and classification of human genome to predict the accuracy of disease effecting a person through machine learning approaches became enormous in the present-day. Research in DNA is unearthing the evolution of human in the nature by mapping the gene based on the physical features. DNA classification of chromosomes can be used for predicting numerous inheritance mutations and DNA damages. Human-Genome-Project which was started in 1990 and continued till 2003 worked on the base's sequences in the gene sections of 3 billion long with an average size of 3000 bases. Human-genome-project work done by scientist all over the world improved the medical field, microbial genome research, forensics, disease risk assessment and human evolution. DNA abnormalities can lay a step forward for predicting the multifactorial inheritance risk in effecting children from parents.

III. RELATED WORK

The classification of eukaryotic genome linear DNA chromosomes on extrachromosomal DNA by Zhenyu liao et al., in 2020 [1] provided a circular structure of the extra chromosomes which were found outside of the eukaryotic-genome. The authors had classified the unconventional chromosomes for identifying the cancer tumor miscellaneous behavior in the gene. The extrachromosomal were classified based of the spectrum of microscopy technology for detecting the progression of tumor. According to the authors research the extrachromosomal DNA is mainly characterized into four types based on the size of the chromosomes, frequency in the tumor cells and functionality. The progression in the tumor were relatedly close to two gene amplification in the chromosome segmentation which increases the intensity in the

tumor cells. The authors discussed about the drugs that resist to the tumor cells and formulated the extrachromosomal cycle and its proliferation in the genome. The extrachromosomal DNA formulation gives the complete information regarding the circular chromosome amplification and its translocation of tumor cells. The super resolution of the next generation genome sequencing for different elements of the extrachromosomal DNA tumor identification and amplification were resolved using gene editing tools.

Receiver-operating-Characteristic-curve was achieved by Leif E.Peterson et al.,[2] for cancer microarray DNA using different Machine-learning classifiers, feature-scaling and fuzzification for the 9 cancer microarray datasets. The sample size is initially considered for obtaining the AUC values for the fuzzy set and the crisp set based on the statistics to recommend the factor which is influencing the AUC percentage value. The inferential hypothesis test is used for predicting the effect on the AUC value. The feature-scaling of the dataset is performed using the t-test suboptimal ranking for N inputs. The fuzzy logics were used to get the real gene expressed in the cancer microarray dataset. Machine learning logics were used to find the regression and classification of the cancer microarray datasets using certain formulas for supervised classification. The authors had compared the results of both the fuzzy and crisp accuracy of the 9 datasets in the graphical representation and the cancer microarray are categorized based on the classification of the data. The AUC fitting for the cancer datasets for the least and the highest correlations are obtained based on the feature-scaling and fuzzification performance.

Stephen winters-hilt et al.,[3] had developed a new computational method for Single-Molecular DNA Classification to enhance the accuracy of hairpin DNA using single species data in silico. This computation method was named as Watson-crick-basepairs. The authors in this model, performed the SVM multiclass architecture to analyze the state of DNA molecule and also its transitions in the molecular structure with respect to the kernels. The Hidden-Markov-model parameters are also used for denoising and feature-vectors in the molecule. The model performance is measured for a single DNA molecule using the performance metrics in the biophysical analysis. Nanopores are generally used to measure the DNA molecules for each and every basepairs, the feature extractions are performed from multiclass scalability trends to analyze the sequential data of basepairs. SVM provides the optimized hyperplane which separates the hyperplane into clusters for mapping of feature-vector to discriminate the structural risk in DNA hairpin in the experimental procedure. This model achieved a highest accuracy of 99.6% in less than six seconds.

A new method for handling classification problem in the DNA coding was proposed by Ting-Cheng et al., in 2015 [4] as variable-coded-hierarchical-fuzzy-classification-model (VCHFM). The supervised learning method is used as an interface for fuzzy system and the DNA coding. This model works on four main principles. The first main principle of VCHFM is automatic fuzzy rules generation for numeric data and feature-extraction. The second principle works on the DNA computation functions. The third principle works on the

optimization of the chaotic particle that regulates the weight grade of the inference node in DNA. The final principle works on classification functions and the multi-objective-fitness optimization function. This model is highly capable of reducing the overlapping problem and dimensionality reduction problem that affects the classification. VCHFM obtained a benchmark result in best classification rate with a smaller number of fuzzy rules.

Umit Atila et al., in 2020 [5] had classified DNA damage problem using convolutional neural network using the comet images of the grayscale DNA. The authors had worked on the quantification of the images and the identification of the damage comet object in the DNA. They divided the entire DNA into four categories namely healthy, poorly defective, defective and very defective based on this the images were classified in the neural network. Comet-assay-experiment was conducted to obtain the images of 170*170-pixel resolution images for four categories labeled as G0, G1, G2 & G3. Authors had achieved a highest accuracy of 96.1% in predicting the damage DNA using convolutional neural network.

R. Touati et al., in 2021[6] provide a detail description on converting the DNA sequences into the chaos-game representation using the FCGR images. Authors had initially considered the helitron-family FCGR images for feature extraction in the automated system to develop the DNA sequences of helitron-family. The authors also applied the machine learning methods for classifying the images of DNA using SVM, PTDNN and Random-forest algorithms. SVM is used to minimize the structural risk in the DNA by dividing it into different clusters in the hyperplane. Pre-trained deep neural network is used for classifying the DNA images using the softmax activation function on the 2D images to classify the images. Random forest techniques like bagging are used for detecting the different variation of DNA images. The accuracy of classification of the DNA with all the three machine learning approaches were analyzed.

L. Liu et al., [7] provided the detailed description of the cancer plasma cell detection from the methylation sequence and its classification. Authors had developed a comprehensive-methylation-sequence by targeting a single plasma and identified the presence of cancer. The molecular testing method is used for classification of the cell-free DNA in the cancer gene to detect the defected plasma.

Sergio Bittanti Simone Garatti Diego Liberati[8] had provided a detailed description of the degeneration effect in the various types of cancer using unsupervised clustering. Author had classified leukemia and paradigmatic using data mining techniques. To analyze the data the authors had used the microarray technology for training the gene expressed data and for performing unsupervised clustering for diagnostic of DNA. Using this approach authors are capable to classify the data without any pathological information.

SNP-Single nucleotide polymorphism impact on DNA is classified by the Jard H. de Vries et al., 2021[9] which is used to solve the investigation of a murder case from the genomic data. The SNPs are generally used for obtaining the DNA quality & quantity, in the crime case by applying the global-

screening-array to the sample. The impact of SNPs is used to classify the kinship, based on the positive and negative values of kinship-classification the murder case is solved.

Jun Hu et al., in 2020 [10] provided information related to the protein sequence analysis from DNA-binding using the computational methods. Authors had worked on the target DNA-binding protein by applying four feature extraction operations. The base features of the DNA are extracted using the Amino-Acid-Composition, Pseudo-position-specific-scoring-matrix, pseudo-predicted-relative-solvent-accessibility and pseudo-predicted-probabilities-DNA-binding-sites. They had combined both the base features and their weights to determine the original-super-feature to perform the machine-learning algorithms. Authors had used statistical predictors to improve the accuracy of the DNA-Binding-protein, dataset analysis, feature extraction, multi-view-features and feature selection to identify the DNA-sequence. The results of applied feature selections are measured using the performance metric like ROC, accuracy and CBR ranking.

The cancer classification based on the DNA-mutation patterns were performed by the Lei Wu et al., in 2020 [11]. The amplifications in the tumor cells are identified in the DNA-sequence using the Surface-Enhanced-Raman-Spectroscopy (SERS). The authors had created a free amplification SERS sensor to integrate it into microfluidic-chip with in the DNA-nucleotide mixture for demonstrating the melanoma-cell lines and colorectal cancer. The SERS classification of cancer types using the profiling mutation in the DNA patterns had achieved a benchmark accuracy of above 90%.

The classification of DNA-microarray using advanced algorithms was discussed by Beatriz A. Garro et al., in 2016 [12] for synthesis of mRNA molecules. The authors had diagnosed various diseases to identify the tumor and detect its amplification in the cells. The new algorithms using artificial-neural-network to solve the classification problem in the DNA groups associated for a particular disease in gene expression. The computational models used by the authors to classify are artificial-neural-network, multilayer-perceptron, radial-basis-functions and support-vector-machine. In the feature selection process, the authors had evaluated the accuracy of the model by using four different datasets for a particular disease using ABC algorithm.

Firoz khan et al., in 2020 [13] provided the information related to the digital DNA-sequencing to detect the ransomware using various machine-learning approaches. Software used to implement digital DNA-sequencing is ransomware, it is one of the classes of malicious software which is used to predict various attacks over the internet. Author had developed appropriate ransomware attack flow with machine-learning algorithms called DNAact-Ran (engine) methodology to predict the DNA-sequence. The DNAact-Ran engine is evaluated using the machine-learning performance metrics for analyzing the accuracy and engine effectiveness on real-time dataset.

The DNA-Binding model EL_LSTM was proposed by Jiyun Zhou et al., in 2020 [14] for residue relationship prediction. In this novel approach the authors had mainly

concentrated on the two concepts initially finding the pairwise relationship with long-short-term-memory bigram model and secondly solving the data-imbalance problem in DNA-binding. The authors considered four datasets namely PDNA-224, DBP-123, HOLO-83 and TS-61. The residue data instances are calculated using the sequence length and sub sequence length of the chain with protein sequence. Using the LSTM method, the experiment was conducted and the performance of each and every residue data instances are evaluated for neural-network, random forest, support vector machine and LSTM and accuracies of all algorithms are analyzed for all the four datasets.

Wunsch algorithm was proposed by Amr Ezz El-Din Rashed et al., in 2021 [15]. Biological-sequence-alignment-algorithm problems were addressed by authors using the Wunsch algorithm. In this model the DNA-sequence is considered as the input to the parallel workflow model, all the input sequences are performed with the machine learning models to decode the sequences label and to obtain the results. The traditional sequential model discussed in this paper are based on the sequential workflow consisting of input sequence, initialization of matrix, matrix score, traceback of matrix score and results generation. The Wunsch algorithm is capable of computing and converting the DNA-sequence from alphabet to the decimal or binary representation. The proposed model achieved a benchmark accuracy of 99.70% to prevent overfitting problem in the DNA-sequence.

Transcription-factor-binding was implemented using convolutional-neural-network by Qinhu Zhang et al., in 2021[16] for understanding the various DNA cellular-functions and binding mechanisms.

Erfan Aref-Eshghi et al., in 2018 [17] provided a detail description on monotonous debate between histone adjustment DNA methylation proposes that the alimnt might be anticipated to show DNA-methylation impressions that ponder those primitive error related with chromatin myopathies. Here we study 14 mendelian states that show from direct disordering or indirect disordering of the proteins. To recognize Genomic regions, bear methylation. Switch, a jolt coursing approach is used by the jolt courser package. Ten-fold cross checking of this representation showed an accuracy of 99.6%. If exactly detected the class of the 141 pompous subject that are used in training with other samples obtainable from alike conditions & other diseases are being reviewed for discovery of epi-signatures.

Antonino Fiannaca et al., in 2015 [18] suggested adjustment free procedure for DNA barcode categorization that is established on both a phantom representation and a neural-gas-network (NGN), for unsupervised-clustering. Best results can be acquired using adjustment -free approaches established on phantom sequence representation. The main aim of the suggested process that is pondered an addition of our earlier work is the categorization of an unrevealed DNA sequence utilizing the frequency of a little set of k-mers. Here other classifiers reached almost (99%-100%) accuracy, but CT method reached 97%. The efficiency decays to approximately 95% at the family, species, and genus level with suggested method 98-96% accuracy is achieved when for analysing full-

length sequences the score decays to 99-97% compared with suggested method.

Sara Alghunam et.,al. in 2019 [19] report on machine-learning that can be used for categorization, handling each dataset individually and connecting them. Micro-array is one of the fastest growing technologies in genetic research. When SVM and logistics were used they showed 50% and 45% respectively after feature selection process is applied, they showed 75% and 63% respectively. Spark & Weka libraries - when analysed spark libraries showed high accuracy and SVM (in Weka) has exceeded the other classifiers. Comparisons showed that GE data exceed DM, and SVM has given 99.68% efficiency.

Wenbin Liu et al., in 2020 [20] proposed a new miscellaneous learning review on ASM-SNP data (bi-polar disorder & schizophrenia). The recognized genetic differences in ASM-SNP data are key to disclose the underneath process of mental problems. Here the authors labelled the immediate confront via the latest miscellaneous learning and machine learning. New SNP (feature-selection) and continuous pathway-selection was examined and various miscellaneous learning meths included kernel-PCA, LLE...etc. The comparison from constitution clustering and imaging suggested that the misclassification between schizophrenia & Bi-polar disorder could be unavoidable for physicians. They achieved highest performance when using t-SNE and needed only 20% SNPs (Top-ranked) to achieve the best diagnosis.

Giulio Pavesi et al., [21] investigating whether practical details about genes can be predicted by using details obtained from their sequences combined with gene expression data. The SVM plays a crucial role in responding to the main request emerged from the work. To evaluate the categorization execution of SVMs authors worked on various holdout mechanism. They cautiously reviewed the election bias issue. For every training set obtained from proponent a subgroup of motifs outcome calculated by weeder. In specific, regarding one cluster, the genes percentage in test sets ranged from 68% to 70%. Experimentation as well case studies such as these may assist to shack more light on the present problem which remains amongst the most pertinent and calculated in bioinformatics and molecular-biology.

DNA functionality method that used the p53 malfunctioning in identifying the diseases [27] was proposed by Mikael S.Lindstrom et al., in 2022. The recognition of p53 worked was focused from many years with the enhancement in the biomedicine. Malfunctions associated with each disease can be identified using the clear insights of p53 malignancies. Authors used p53 for treating the cancer patients based of the DNA replicas. The fundamental procedure used for treating the cancer patients are p53-centeres multifaceted pathways and ribosome-biogenesis (RiBi). Author using the DNA replica and RiBi methods proposed a new approach which firstly deals with p53 canonical interaction and their regulation in post-translation of the target. Secondly, the response in DNA speed in performing replication with the p53 cellular genomic association with targeted cancer cells are given brief description. Emerging of p53 in replication stress (RS) are highlighted from the emerge of p53 to the key role in DNA

link replication. In addition, the tantalizing crosstalk in identifying the mediated monitoring between DNA replication and cell nucleolar RiBi are analyzed. Cancer diseases are outlined using the IRBC and the RiBi pathway tumorigenesis identification using the p53 malfunctioning in human. The p53 role in identifying the DNA replication and ribosome-biogenesis in cell homeostasis provided a clear vulnerability in identifying the cancer elucidation.

Prognostic investigation on DNA [28] methylation to identify the subtypes of tumors are proposed by Christopher et al., in 2022. Aberrant analyses of human DNA methyl patterns helped in identifying the subtypes of cancer diseases based on the response and outcomes. Authors used osteosarcoma malignancy procedure to perform chemotherapy using DNA methylation analyses. Authors worked on predicting the patient tumor with the help of the genomic methylation to identify the situation in the early stages. The patient response behavior to surgical reactions is also predicted using the hypomethylation procedure which derived high perfect outcomes. Downstream analysis for identifying the methylation patterns were performed in an experimental analysis using three datasets to derive site-specific methyl patterns. The experimental analysis was associated with the clinical human genomic outcomes.

Impact of mitochondrial DNA (mtDNA) in human brain postmortem detailed description was provided by Alba Valiente-palleja et al., in 2022. The authors [29] investigation pm mitochondrial DNA reveals the facts on the heterogeneous disorder genes that synthesize the phosphorylation oxidative systems. Neuropsychiatric symptoms are used for understanding the disorders of the human brain functioning. authors provided an empirical study on human brain tissues to alert the ageing process investigation in unequivocally diseases. The experimental analysis of this procedure on testing with various samples resulted a benchmark outcome in identifying the disorder using the mtDNA for finding the contradictory cells in human brain

IV. ML APPROACHES

In order to process the genetic information, machine learning algorithms plays a vital role in building a computational model by using the statistical theory. Table 1 describes the accuracy of various machine learning algorithms applied on different genome datasets. It helps the researchers in analyzing the DNA and its replications to various diseases and ML helps to identify the abnormalities in using the experimental procedure and optimization techniques to derive the accurate outcomes.

TABLE I. ACCURACY OF VARIOUS MACHINE LEARNING ALGORITHMS APPLIED ON DIFFERENT GENOME DATASETS

DATASET	ALGORITHM	ACCURACY
Helitrons database	Pre-Trained Deep Neural Network (PTDNN) classifier	72.6%
Helitrons database	Support vector machine (SVM)	68.7%
Helitrons database	Random forest (RF)	91%
TF Binding Datasets	Deeper CNN	

UCI Pima Indians Diabetes	Fuzzy rule-based classification	73.70%
Glass	Fuzzy rule-based classification	60.04%
Wisconsin Breast Cancer,	Fuzzy rule-based classification	91.21%
Wine	fuzzy if-then rules.	99%
Iris datasets	Fuzzy rule-based classification	96.67%
PDB database	SVM-REF+CBR; without feature extraction	78.85
PDB database	SVM-REF+CBR; with feature extraction	79.71
Multiple datasets	Needleman–Wunsch (NW) algorithm	85.9
WDBC directory	Sequential minimal optimisation (SMO),k-nearest neighbour(KNN),and decision tree(BF-tree)	96.19%
WDBC directory	Hybrid of k-means and SVM	97.38%
The gene expression omnibus(GEO)	novel graph-based semi supervised learning algorithm	24.9%
National center of biotechnology information (NCBIGEO)	SVM and logistic regression	>75%
Orange laboratories	SVM and logistic regression on sparks	75%
Epsilon dataset and GECCO dataset	SVM logistic regression, and Naive Bayes on spark	>75%
NCBI GEO	SVM	<70%
PDNA-224	EL_LSTM	82.59
DBP-123	EL_LSTM	81.44
PDNA-224	LSTM	78.36
DBP-123	LSTM	80.51
PDNA-224	NN	72.34
PDNA-224	Rf	75.27
PDNA-224	SVM	74.98
PDNA-224	LSTM	78.36
DBP-123	NN	76.36
DBP-123	Rf	77.29
DBP-123	SVM	78.34
DBP-123	LSTM	80.51
Ransomware	Multi-Objective Grey Wolf Optimization (MOGWO)	78.5%
Ransomware	Binary Cuckoo Search (BCS) algorithms.	83.2%
ALL-AML	MLP	1.0000
ALL-AML	SVM	1.0000
ALL-AML	SVM	1.0000
ALL-AML	KNN	0.9736
ALL-AML	SMV	0.9583
ALL-AML	KNN	0.9412
BREAST	SVM	1.0000
BREAST	SVM	0.9470

BREAST	J48	0.9381
BREAST	SMV	0.8421
PROSTATE	MLP	1.000
PROSTATE	SVM	0.9804
PROSTATE	SMV	0.9706
PROSTATE	LDA	0.9550
PROSTATE	LDA	0.9118
BOLD database	SVM	64.8%
Comet assay database	CNN	96.1%.

V. CURRENT CHALLENGES

The researchers, who are doing research with the DNA dataset by applying the machine learning logics can work on this current challenges.

- Improvements in the medical field.
- SNP for forensic investigation.
- DNA samples for identifying the human genetics.
- Inheritance disease identification.
- DNA damage in the gene sequence.
- DNA sequencing accuracy enhancement.
- DNA degradation.
- DNA analysis for body fragmentation.
- DNA mutation identification.
- DNA affect in health development.
- DNA profiling.
- DNA analysis with the high-end technologies.

VI. CONCLUSION

This paper helps the researcher those who are doing research with the DNA dataset to choose the appropriate machine learning logics depending on the model accuracy. They are given scope for predicting the sequence in the genome for extrachromosomal amplification identification using the feature scaling. By applying machine learning techniques in the unconventional chromosome can minimize the structural risk. DNA mutations can also be predicted using the genome data analysis using the statistical theory. Experimentation as well case studies shown in this paper may assist to shack more light on the present challenges.

REFERENCES

[1] Liao, Zhenyu, et al. "Classification of extrachromosomal circular DNA with a focus on the role of extrachromosomal DNA (ecDNA) in tumor heterogeneity and progression." *Biochimica et Biophysica Acta (BBA)-Reviews on Cancer*, 2020.

[2] Peterson, Leif E., and Matthew A. Coleman. "Machine learning-based receiver operating characteristic (ROC) curves for crisp and fuzzy classification of DNA microarrays in cancer research." *International Journal of Approximate Reasoning*, vol.47, pp. 17-36, 2008.

[3] Winters-Hilt, Stephen, et al. "Highly accurate classification of Watson-Crick basepairs on termini of single DNA molecules." *Biophysical Journal*, vol.84, pp. 967-976, 2003.

[4] Feng, Ting-Cheng, Tzoo-Hseng S. Li, and Ping-Huan Kuo. "Variable coded hierarchical fuzzy classification model using DNA coding and evolutionary programming." *Applied Mathematical Modelling* vol.39, pp.23-24 (2015): 7401-7419.

[5] Atila, Ümit, et al. "Classification of DNA damages on segmented comet assay images using convolutional neural network." *Computer methods and programs in biomedicine* vol.186 (2020): 105192.

[6] Touati, R., et al. "New intraclass helitrons classification using DNA-image sequences and machine learning approaches." *IRBM* vol.42.3 (2021): pp.154-164.

[7] Liu, L., et al. "Targeted methylation sequencing of plasma cell-free DNA for cancer detection and classification." *Annals of Oncology* vol.29.6 (2018): pp.1445-1453.

[8] Liberatib, Sergio Bittantia Simone Garattia Diego. "From DNA Micro-Arrays to Disease Classification: an Unsupervised Clustering Approach." (2005).

[9] de Vries, Jard H., et al. "Impact of SNP microarray analysis of compromised DNA on kinship classification success in the context of investigative genetic genealogy." *Forensic Science International: Genetics* vol.356 (2022): 102625.

[10] Hu, Jun, et al. "TargetDBP: accurate DNA-binding protein prediction via sequence-based multi-view feature learning." *IEEE/ACM transactions on computational biology and bioinformatics* vol.17.4 (2019): pp.1419-1429.

[11] Wu, Lei, et al. "Profiling DNA mutation patterns by SERS fingerprinting for supervised cancer classification." *Biosensors and Bioelectronics* 165 (2020): 112392.

[12] Garro, Beatriz A., Katya Rodríguez, and Roberto A. Vázquez. "Classification of DNA microarrays using artificial neural networks and ABC algorithm." *Applied Soft Computing* 38 (2016): pp.548-560.

[13] Khan, Firoz, et al. "A digital DNA sequencing engine for ransomware detection using machine learning." *IEEE Access* 8 (2020): pp.119710-119719.

[14] Zhou, Jiyun, et al. "EL_LSTM: prediction of DNA-binding residue from protein sequence by combining long short-term memory and ensemble learning." *IEEE/ACM transactions on computational biology and bioinformatics* 17.1 (2018): pp.124-135.

[15] Rashed, Amr Ezz El-Din, et al. "Sequence Alignment Using Machine Learning-Based Needleman-Wunsch Algorithm." *IEEE Access* 9 (2021): pp.109522-109535.

[16] Zhang, Qinhu, Zhen Shen, and De-Shuang Huang. "Predicting in-vitro transcription factor binding sites using DNA sequence+shape." *IEEE/ACM transactions on computational biology and bioinformatics* (2019).

[17] Aref-Eshghi, Erfan, et al. "Genomic DNA methylation signatures enable concurrent diagnosis and clinical genetic variant classification in neurodevelopmental syndromes." *The American Journal of Human Genetics* 102.1 (2018): pp.156-174.

[18] Fiannaca, Antonino, et al. "A k-mer-based barcode DNA classification methodology based on spectral representation and a neural gas network." *Artificial intelligence in medicine* 64.3 (2015): pp.173-184.

[19] Alghunaim, Sara, and Heyam H. Al-Baity. "On the scalability of machine-learning algorithms for breast cancer prediction in big data context." *IEEE Access* 7 (2019): pp.91535-91546.

[20] Liu, Wenbin, Dongdong Li, and Henry Han. "Manifold learning analysis for allele-skewed DNA modification SNPs for psychiatric disorders." *IEEE Access* 8 (2020): pp.33023-33038.

[21] Pavesi, Giulio, and Giorgio Valentini. "Classification of co-expressed genes from DNA regulatory regions." *Information Fusion* vol.10.3 (2009): pp.233-241.

[22] Kaur, Pinderpal, et al. "DNA damage protection: an excellent application of bioactive compounds." *Bioresources and Bioprocessing* vol.6.1 (2019): pp.1-11.

- [23] Shi, Bingyang, et al. "Challenges in DNA delivery and recent advances in multifunctional polymeric DNA delivery systems." *Biomacromolecules* 18.8 (2017): pp.2231-2246.
- [24] National Human Genome Research Institute: <https://www.genome.gov/about-genomics/fact-sheets/Deoxyribonucleic-Acid-Fact-Sheet>.
- [25] Lindahl,T,Wood,RD, Reactome, Quality control by DNA repair: <https://reactome.org/content/detail/R-HSA-73894>.
- [26] Ludovic Bourré, Crown bioscience: <https://blog.crownbio.com/dna-damage-response>.
- [27] Lindström, Mikael S., Jiri Bartek, and Apolinar Maya-Mendoza. "p53 at the crossroad of DNA replication and ribosome biogenesis stress pathways." *Cell Death & Differentiation* (2022): 1-11.
- [28] Lietz, Christopher E., Erik T. Newman, Andrew D. Kelly, David H. Xiang, Ziyang Zhang, Caroline A. Luscko, Santiago A. Lozano-Calderon et al. "Genome-wide DNA methylation patterns reveal clinically relevant predictive and prognostic subtypes in human osteosarcoma." *Communications biology* 5, no. 1 (2022): 1-20.
- [29] Valiente-Pallejà, A., Tortajada, J., Bulduk, B. K., Vilella, E., Garrabou, G., Muntané, G., & Martorell, L. (2022). Comprehensive summary of mitochondrial DNA alterations in the postmortem human brain: A systematic review. *EBioMedicine*, 76, 103815.

An Architecture of Domain Independent and Extensible Intelligent Tutoring System based on Concept Dependencies and Subject Paths

Sanjay Singh, Vikram Singh

Department of Computer Science and Engineering
Chaudhary Devi Lal University, Sirsa, India

Abstract—Intelligent Tutoring Systems (ITS) seek to provide personalized tutoring to learners, but are often domain specific, and lack extensibility. When featuring extensibility and domain independence, it is a challenge to provide appropriate level of personalization for every learner. In this paper, an architecture of a system that features domain-independence and extensibility with personalization and automatic course improvements without requiring persistent subject expert intervention has been proposed. The proposed architecture utilizes the notion of concept dependencies and the ability to sequence inter-dependent concepts intelligently into subject paths that enable automated tutoring as well as effective course customization per learner. It features a separate interface for subject experts through which they do not require ITS building knowledge to fulfil their appropriately assigned tasks assisted intelligently by the system, and an API based interface layer that supports today's mobile requirements for better engagement.

Keywords—Personalized tutoring; intelligent tutoring system; adaptive learning

I. INTRODUCTION

Since centuries, students have been taught in classrooms, where there is one teacher and multiple students. A more personalized tuition may involve a teacher personally tutoring a single student. Ever since computers were invented, there has been a considerable effort at using them to mimic the teaching capabilities of a human teacher. Facilitating the job of teaching through the use of any electronic technology falls under the huge umbrella of e-learning, which is a far-reaching discipline that covers the analysis of all conjunctions of technology and education. A more appropriate definition by researchers in [1] states “Educational technology is the study and ethical practice of facilitating learning and improving performance by creating, using, and managing appropriate technological processes and resources”.

Since a human teacher, besides barely delivering a classroom lecture, performs considerable communication with the students and adjusts their teaching in response to the students' learning progress, researchers have been aiming to bring the exact qualities into e-learning systems to provide better tutoring. The most important and in-fact the ideal characteristic for an intelligent tutor is the ability of this communication [2]. The years around the 1960s and 1970s saw many new Computer-Assisted Instruction (CAI) projects funded by big names such as IBM and HP, that looked at

tutoring through a behaviorist perspective based on Skinner's theories [3]. While the CAIs were attracting interests, the researchers in [4] introduced Intelligent Tutoring Systems (ITS) and shed light on the idea that computers could act as a teacher rather than barely a tool. These systems realize established teaching-learning processes by way of AI (Artificial Intelligence) with an intention of delivering learner-adapted tutoring without any direct mediation of a human teacher. Intelligent tutoring systems combine AI, education theories, and psychological models of the student and the expert [5]. Thus, building truly intelligent tutoring systems requires experts from the AI community, psychology community and education community to come together. In a nutshell, an ITS aims to put AI technologies to use for the delivery of a teaching, which would have been branded as “Good Teaching” if it were delivered by a human teacher [6]. Many of the works have employed ethnographic and design research methods [7] to study the scenarios in which the teachers and learners actually use the ITSs, many a times disclosing unexpected needs that they met, failed to meet or even create in some cases.

Present day ITSs try to mimic the role of a teaching assistant, by trying to automate pedagogical functions such as problem generation and selection. Many recent works have focused on how ITSs can supplement the duties of an existing human teacher [8] in a classroom or a peer [9] in other social contexts. The field of intelligent tutoring systems have evolved into various sub-areas, like Dialogue Based Tutoring Systems [10] which provide tutoring to the student using natural language dialogue, Cognitive Tutors [11] which utilize a cognitive model (an approximation of animal cognitive processes) to provide feedback during the learning process, tutoring systems for Intelligent Computer Assisted Language Learning, etc.

The task of individualized one-to-one tutoring can be made more efficient if the tutoring machine is portable, leading us to the field of mobile learning, which allows a student to carry the tutor with them wherever they go. According to [12] M-learning is “learning across multiple contexts, through social and content interactions, using personal electronic devices”. The researchers in [13] suggest that mobile technology combined with network technology can connect formal learning to informal learning. Formal learning refers to the education delivered in a traditional classroom environment by trained teachers and informal learning is any type of learning

which is not formal. Mobile learning has the strength of being portable, which leads to more responsiveness and the ability to provide instant feedback. The work of [14] observed that mobile learning can increase exam scores from the 50th to the 70th percentile and reduce the dropout rate by 22 percent in technical fields. Mobile learning is also relatively cheaper because the cost of mobile devices is relatively lesser than that of laptops and personal computers.

Intelligent tutoring systems are expensive to develop, are often built for a specific domain of study, and the domain knowledge contained in them is limited because it is usually only fed-in at the time of development of the system. This paper proposes a framework for developing a domain independent and extensible system, that does not require persistent availability of subject experts, and the following sections discuss the related work, the proposed system's architecture with description of its components and modules, and an evaluation of the modules of the system.

II. RELATED WORK

The architecture of an ITS has greatly varied throughout the years in the works of various researchers but the core idea has remained the same – it can employ any architecture provided that it delivers intelligent tutoring, even though the original discussions on the architecture of an ITS describe these four crucial modules: The Student Model, The Domain Model, The Tutor (pedagogical) Model and The Interface Model [15–17]. The domain model, which can also be referred to as the knowledge model is the part of the system that contains the concepts, rules and problem-solving strategies of the domain to be learned. It helps fulfilling certain roles such as – being a source of knowledge, being a standard for evaluating learner's performance, etc. The student model is generally built on top of the domain model, as an overlay. It is often referred to as the core component of any ITS, as it deals closely with the learner's cognitive and affective states and their evolution as the learning process advances. The tutor (or pedagogical) model connects with the domain and student model of the system and makes decisions about the tutoring actions and strategies. This model is responsible for guiding the learner through the overall learning process making sure the learner does not deviate from the particular tutoring strategy adopted by the system. The interface model “integrates three types of information that are needed in carrying out a dialogue: knowledge about patterns of interpretation (to understand a speaker) and action (to generate utterances) within dialogues; domain knowledge needed for communicating content; and knowledge needed for communicating intent”

It is in-fact extremely rare to find two different scratch ITSs with the same architecture. The word Scratch ITS refers to all those ITSs that have been built from the ground up and have not been authored using some ITS building framework. Out of these, there have been ITSs based on three-model, four-model and other varieties of architecture.

A three-model architecture is based upon the declaration of three major core components – domain knowledge, student knowledge and tutoring knowledge. Researchers in [18] have proposed an ITS with the expert domain model, tutoring model and student knowledge model being the three major

components. The expert domain model and the student knowledge model guide the procedures in the tutoring model by providing the necessary information. The tutoring model is the most well-defined part of this architecture and it has various sub-components for curriculum planning, tutorial intervention and lesson planning. Another important three-model architecture is in the work of [19]. It also comprises of a domain model, a tutoring model and a student model but the difference between this architecture and the previous one is that it also incorporates an additional process – an overall system control process to co-ordinate the three models. This architecture also extends the lesson planning and dynamic adaptation concepts from the previous architecture to facilitate multiple tutoring strategies and information representations.

The three-model architectures made way for the classical standard four-model architectures. These architectures contain the three core components discussed in the previous architectures and add a fourth – the user interface component. A typical example for the four-model ITS architecture would be the work in [20] that has a knowledge base, a student model, a pedagogical model and a user interface. This architecture embodies cognitive and meta-cognitive processes in the student model and contains domain dependent tutoring rules in the pedagogical model. The key difference between this architecture and the architectures discussed in the previous section is that this architecture regards the user interface as an integral and internal component of the system whereas the previous section regarded this as a component external to the whole ITS system. This inclusion is helpful in that it concretizes the fact that a user-interface has a huge impact on the overall tutoring process, hence more efforts in user-interface design and development have become a concerning part of the overall ITS development process.

There are some architectures that take a deviance from the three-model and four-model architectures discussed. The idea of an intelligent learning environment has been promoted in MATHEMA [21] a multi-agent architecture which incorporates the notion of a human expert society (HES), a micro-society of artificial tutoring agents (MARTA) and external human motivators. This architecture surrounds the constituents of the classical architectures, although through a distinctive representation. The domain model is implanted within HES and MARTA, the user interface component is represented as interface agent, the role and functions of the tutoring model are dispensed among MARTA, and the human learner is not represented as a student model component as in other architectures, it is instead represented as a component of the learning environment itself. This architecture is suggested for well-structured, formal and specific knowledge domains.

There has been very limited work that has been done in the direction of domain independent ITSs, for they require considerable effort in representing each domain fairly. The researchers in [22] put forward ASSISTment builder tool that allows easy creation of ITSs that mimic cognitive tutors but the ITSs made with this are limited in that they only work for a single problem. These tutors provide a simple cognitive model based upon a state-graph tailored to a specific problem. Other aspects that have been rarely touched are extensibility – the ability to considerably extend the knowledge contained in the

system even after the system has been deployed. AutoTutor [23] is an intelligent tutoring system that helps students learn science, technology, and other technical subject matters by holding conversations with the student in natural language. Being able to extend the domain knowledge easily without requiring expert ITS development knowledge and doing that on-the-fly still remains a challenge. Another challenge is portability – for having a greater amount of time of the day with the learner, which brings various opportunities for better tutoring motivation. For solving all these challenges and more, in the following section, an architecture that ensures domain independence, extensibility, and portability, while ensuring personalization and adaptivity has been proposed.

III. CAPTAIN: MODULES, ARCHITECTURE AND WORKFLOW

This paper proposes Computer Assisted Personal Tutor with Adaptive INstruction (CAPTAIN) - a domain-independent ITS based upon concept dependencies and subject paths. It features extensibility – through automatic data generation, learner submitted questions and impersistent expert intervention, and portability – through an API based interface model that allows both desktop and mobile clients to interact with the ITS.

Its main strength is being able to expand its knowledge, improve itself, and be adaptable to any study area. It is based on five-module architecture and the modules have further specifications for sub-components as illustrated in detail in Fig. 1.

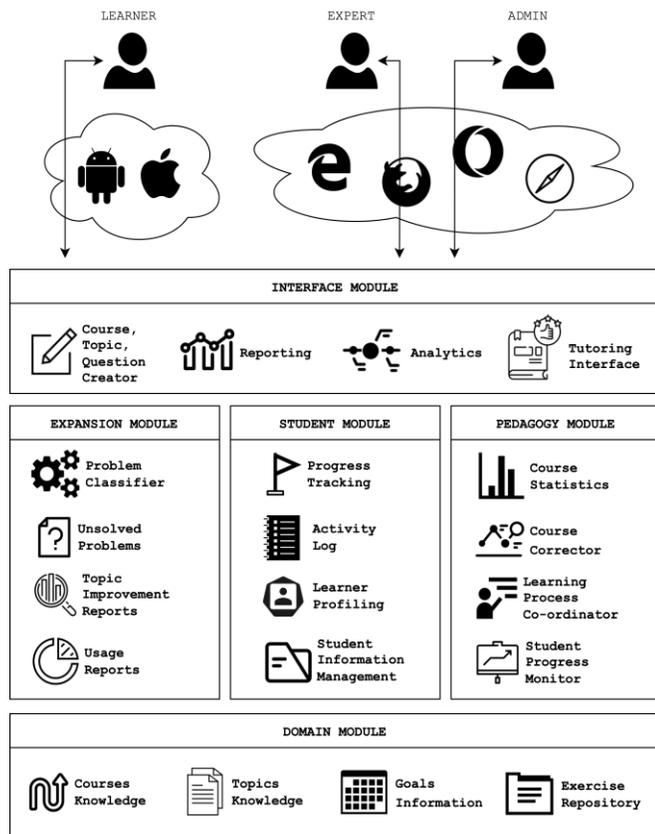


Fig. 1. Architecture Diagram of the System.

The domain module stores the domain knowledge. The architecture is not designed for any specific study domain, so the domain module is capable of storing knowledge about almost any subject area that can be represented through storing and linking modular topics, courses and study paths, along with exercises represented as questions and activities, and information about goals that drive the study paths and the whole tutoring flow.

The pedagogy module is made up of different components that share the common goal of assisting the tutoring process. It is responsible for using learner activity data for improving the quality of the courses that have been crafted by subject experts through its course corrector component, as well as coordinating the learning process and monitoring the progress of the learner.

The student module deals with storing and managing all the information about the learner. It deals with learner profiling – customizing the course according to the knowledge level of the learner in the course, and also deals with tracking the progress and logging the activities of the learner which are in turn used to by various other components of the system.

The expansion module is responsible for the ever-growing nature of the system. It has the problem classifier which maps the user submitted problems to their closest subject area for effective resolution by the appropriate subject expert, or prompts for creation of new topics if there aren't any appropriate ones available. It also provides the subject experts with topic improvement reports and information about how the courses are performing as well as the unsolved problems. It also provides the administrators the comprehensive usage reports of the system.

The interface module is responsible for facilitating the interaction between the learner and the system through mobile based platforms, and between the subject expert and the system as well as between the administrators/management and the system through the web-based platforms.

The system is designed and developed and has been deployed as three interconnected components – a backend server based on Django (python) that uses PostgreSQL as the database engine and serves a RESTful API, a web-app primarily for subject experts and administrators based on React JS, and a mobile app for the learners based on React Native.

In the following sections the key parts of the proposed architecture are explained in detail.

A. Basis

At the core of the system, the smallest and most re-usable atomic unit of information that exists to teach a fine-grained concept or skill has been termed a topic. Any subject area that is intended to be tutored must be fed into the system by subject experts in the form of small topics. The subject experts are expected to create topics that consist of that atomic bit of information which is content in itself.

The amount of information on a concept which is small enough that it is always expected to be studied together in one go, and that dividing which further will not make enough sense as there will always be another bit of information that always needs to be studied along with it. This makes the topics re-

usable and they can be used again and again in different courses, which in turn can be re-used in different study paths. As depicted in Fig. 2, in the proposed architecture, a topic is made up of any sequence of theory activities and question activities, along with an evaluation activity.

As it stated above that a topic needs to be fine grained, it may naturally require prior knowledge of zero or more other topics, which is necessary to represent complex concepts as simple learnable units. Topics are related with other topics via the ‘requires’ relation. This relationship can be realized via a graph, more precisely through a Directed Acyclic Graph representing topics as nodes and a topic’s requirements as an edge from the topic’s node to its required topic’s node as illustrated in Fig. 3.

Since specifying a topic as a requirement for another topic is susceptible to possible creation of cyclic requirements (a situation in which topic A requires topic B and, directly or indirectly, topic B requires topic A), a measure has been implemented to prevent generation of cycles at the point where the subject expert adds a topic as a requirement for a topic.

Studying individual topics is good, but to study a subject or a large topic, one has to study a large number of topics in a certain sequence. This feature has been termed as a **course**, which is crafted by a subject expert by linking a set of topics together in a particular sequence so as to fulfill the aim of tutoring a subject or a large or complex topic. Since a topic may require prior knowledge of other topics, this puts some constraints on the sequencing of topics in a course, because a topic’s required topics must only be sequenced before the topic. Also, how many of the topics’ required topics will be included in the course is another decision the subject expert makes. If a course contains topics whose pre-required topics are not a part of the course, they become the requirements of the course itself, as illustrated in Fig. 4, with reference to the topic relationships illustrated in Fig. 3, as topics 9 and 7 require topic 2 and topics 10 and 7 require topic 6 but topic 2 and 6 are not part of the course, they become the requirements of the course, rest all other topics’ requirements are being satisfied within the course.

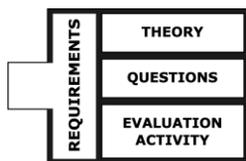


Fig. 2. Structure of a Topic in the System.

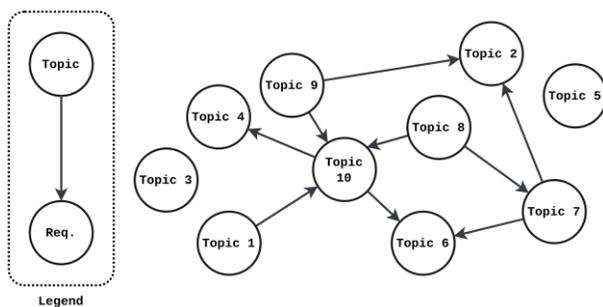


Fig. 3. Topic Relationships Illustrated through a DAG.

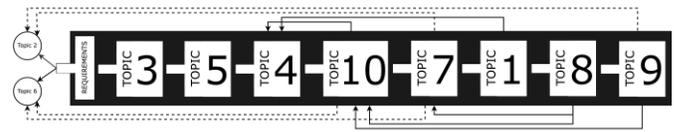


Fig. 4. Illustration of the Structure of a Course.

A course is also designed with such a level of granularity that it is independent in itself in teaching a subject or a large topic, and in this sense, it becomes a re-usable unit that can be studied by learners who have different goals but they are to study certain courses, in a certain sequence to achieve their specific goal. This level of sequencing has been termed a study path, which is an ordered collection of courses intended to prepare the learner for a specific goal, and this is the largest learnable unit in the system, and the least re-usable one.

In the system a goal refers to any concrete motive or objective for studying. The best example of a goal is any national level exam that is to be conducted, and there are many students that need to prepare for it. Goals are added by system administrators, the subject experts then create the study paths for them by re-using the courses in the system or by creating new courses, by in-turn re-using the topics in the system or by creating new topics, while mapping dependency relations to other topics if necessary. Fig. 5 illustrates the operation from the learner’s goal selection to the system’s tutoring process.

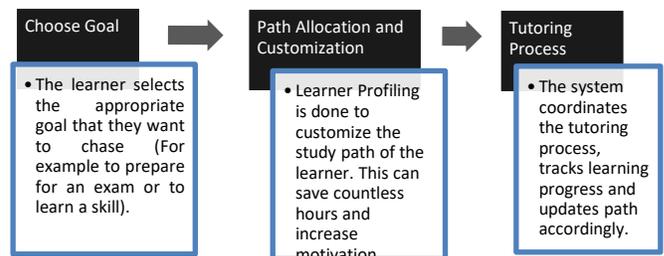


Fig. 5. Flow from Goal Selection to Tutoring.

B. Extensibility

The limited domain knowledge contained in a system can be extended from two ends. One of these is when the subject experts manually feed new domain knowledge into the system using the appropriate interface. This has a limitation – it is slow and it only expands the system in the direction which is determined by what the subject experts think should be and should not be a part of the system. Another way to truly steer the direction of expansion of the system towards what the learners of the system actually need is by allowing the learners to submit new questions.

In the proposed architecture, the new questions submitted by any learner are mapped on to the topics that currently exist in the system, and if there are no appropriate topics that fully relate to the questions, the architecture prompts the subject expert to create new topics that map with the question more appropriately. This leads to two outcomes, the first one being improvement and expansion of already existing topics, which is in turn going to help other learners as well. The second one being creation of new topics, which will eventually lead to generation of new courses, and new courses lead to new

subject paths that lead to new goals that increase the overall reach and usability of the system. Fig. 6 depicts the overall process of expansion.

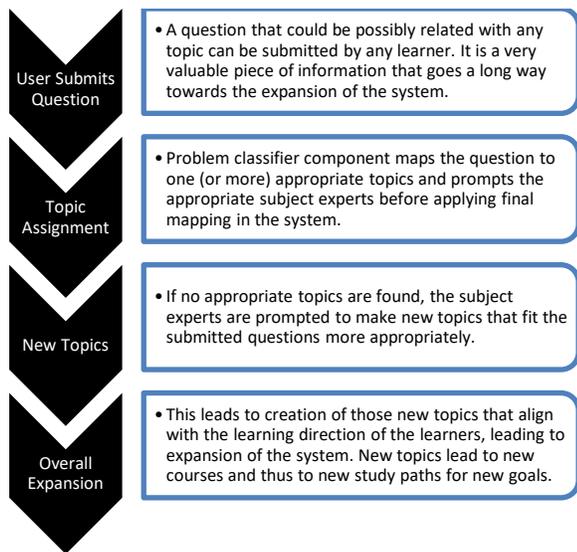


Fig. 6. Flow of System Expansion.

There is a mapping between questions and topics, that a question could be linked to one or more topics. The task of mapping a submitted question/problem to a topic is achieved using the problem classifier component of the system, which is an integral contributor towards the extensibility of the system.

1) *Problem classifier:* The ability of a learner to ask questions to a subject expert for an effective resolution is, at its core, one of the most fundamental activities involved in any learning process. In fact, a good question, combined with its solution will continue to help future learners who want to learn the subject, who can learn a lot by reading other people's questions and their solutions. For this, and many other reasons, in an e-learning environment that deals with questions, it becomes very helpful to classify the questions according to subject areas for an organized study. This problem of question classification falls under the umbrella of a domain of research called text classification.

Text classification deals with utilizing machine learning techniques to assign a class/label to any input text. It has been used widely to classify product reviews by corporations that want to understand the sentiment of their customers, or many specific cases such as to predict the ideological direction of court cases [24], or to classify fake news or hoax, which have become the most prevalent cybercrime in today's day and age that have immeasurable harms [25] or to identify a person's distinctive habits and behaviors through personality classification [26]. Various papers resolve the issue of programmed text classification proposing various strategies and arrangements. Extensive thorough reviews also exist that document text classification in detail [27-30].

When the data-set consists of user-submitted short questions with a vastly disproportionate number of questions in different subjects, the overall classification becomes a

challenge. Researchers in [31] have proposed a general-purpose approach to assigning user-submitted questions their appropriate topic labels, without any extra vocabulary information related to the subjects. Several grid hyperparameter-searched iterations of Generalized Linear Model, Deep Learning (ANN), Gradient Boosting Model, Extreme Gradient Boosting Model (XGBoost), Distributed Random Forests and Extremely Randomized Trees were trained and evaluated, and it was found that out of these, XGBoost performs the best with a small and imbalanced dataset of very short text documents. To further improve the classification performance, a general-purpose approach to handle the unbalanced classes was used utilizing class weights.

C. Personalization

Personalization is the gist of intelligent tutoring. There is no point in calling a system intelligent unless it has some sort of personalization or adaptivity in it. In the context of ITSs, personalization and adaptivity can be practiced at various stages, but the most common one of these is the customization of the course a learner is taking according to the knowledge levels of the learner. This level of adaptivity is helpful at various aspects, one of which is it shortens the overall duration of the course. It is also helpful for uplifting the motivation of the learner to study because if the system presents those topics to the learner which they already know, it can lead to loss of interest, possibly leading to the learner leaving the course in between.

It is particularly challenging to implement course personalization in a system that features extensibility, as the system is going to require personalization in the courses of various varied domains that are going to be created automatically (or subject expert assisted) in the future and do not exist at the time the system was built.

1) *Learner profiling:* To achieve personalization and adaptivity in a system, sufficiently granularized domain knowledge is needed which can be broken-down or combined in multiple ways. This is best achieved when a limited piece of domain knowledge about a certain subject or a fixed number of subjects is structured into the domain model of the system in any rule-based analogy, as utilized by researchers in various domain-specific ITSs discussed in the previous sections. It is easy to form specific strategies to estimate the knowledge level of a specific subject, but to achieve learner profiling in a general-purpose system that supports all types of textual subjects is a challenge that researchers in the field of intelligent tutoring have been trying to solve since decades.

Researchers in [32] have proposed a learner profiling algorithm which is able to adapt any general-purpose course that the learner wants to take, to the knowledge level of the learner, in a very short time quantum, improving the motivation of the learner, shortening the total amount of time needed to complete the course tremendously and hence the overall learning process. The amount of time saved for the learner depends upon the number of topics detected to be known and the individual lengths of those topics. The algorithm requires topic inter-dependency information, which is typically provided by subject experts while drafting the

course. In addition to this, data of responses of past students that undertook the same course is used to improve the topic interdependency information – taking a safer approach for a general-purpose tutoring system when there is a possibility that the subject expert might not be able to map accurate relationships between topics perfectly. These two help the minimal learner profiling algorithm build the learner’s knowledge profile as illustrated in Fig. 7 and it does that in the minimum amount of time possible.

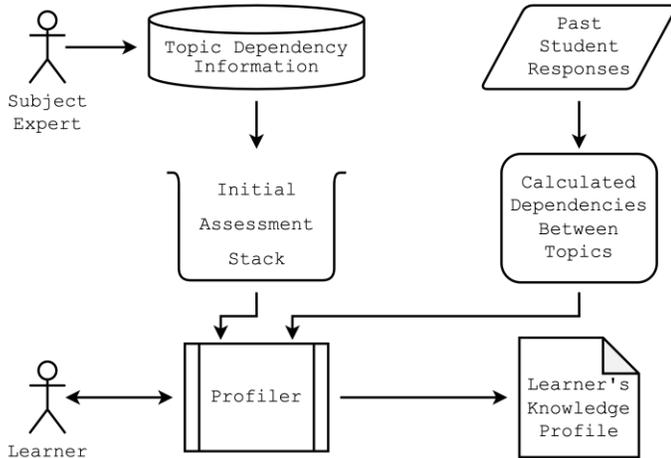


Fig. 7. The Proposed Model for Learner Profiling.

As there is no way to actually look inside the brain of the learner and check the neural connections to determine whether the learner knows the topic or not, the actual accuracy of predicting the knowledge level of the student with respect to a specific topic depends upon the quality of the assessment activity of the topic and the honesty of the learner’s responses. If the quality of the assessment activities is assumed to be perfect the algorithm guarantees accurate adaptation of the course to the learner in the minimum possible time.

D. Course Improvements

A system that allows extensibility and teaches any types of learners without need of persistent intervention from a subject expert, it is necessary to have a mechanism that improves the domain knowledge and course information in the system. In the proposed architecture course improvement is achieved through two means as discussed in the following sections.

1) *Automatic course improvement:* The subject expert specifies the dependencies between topics and creates relationships mappings illustrating which topic depends upon which zero or more topics. These dependencies assist the overall instruction procedure for the learner, but the actual degree of dependencies between the topics is further refined using the data obtained from the initial learner profiling algorithm as covered in the work of [32]. This improves the overall mappings between topics, thus creating even better tutoring procedures and experiences for future students.

2) *Collaborative course Improvement:* This is achieved through manual intervention by the learners. The learner – the ultimate end-users of the system can also, collaboratively,

seed course improvements for everyone. Since the expansion in the system happens dynamically, at any point can have gaps and/or faults in any grain of knowledge anywhere from a simple question to a topic to a course to an entire goal. The learners can report errors or gaps in any grain of the system, and it is prompted to the appropriate subject expert. Since every grain of knowledge is interconnected deeply with all the other parts of the system, improvements in any tiny fragment leads towards refinement of the entire system.

IV. EVALUATION

For the evaluation of the effectiveness of the proposed framework, the components that make up the framework were evaluated thoroughly. The said evaluations have been stated in this section grouped under the respective modules they serve.

A. Expansion Module

In the implementation of the system for the evaluation of the module which is responsible for the ever-growing nature of the system, majorly because of the problem classifier component, a dataset of 6925 problems made up of short text documents from 13 different topics as illustrated in Fig. 8 was taken.

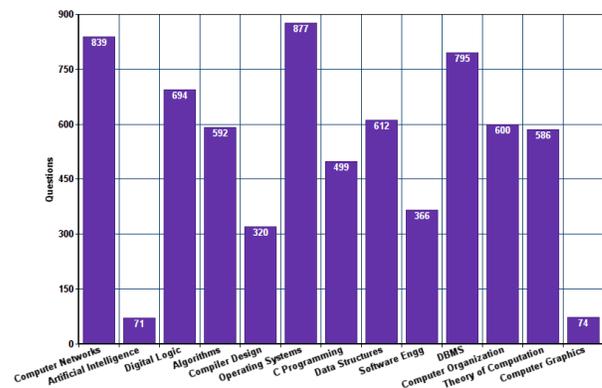


Fig. 8. Problem Dataset Distribution.

The system trained and evaluated around 20 industry standard classification algorithms on the dataset, the four best performing ones of which are shown in Fig. 9 with their per-class error distributions.

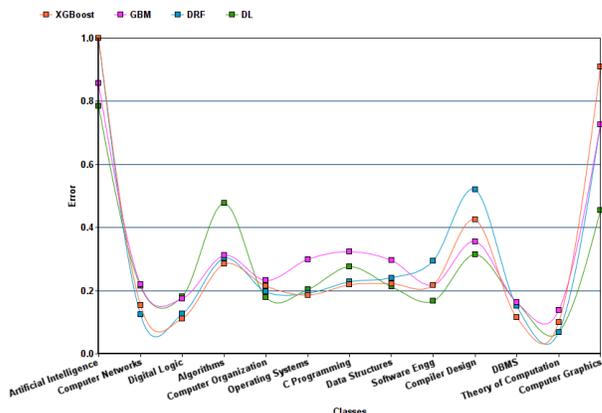


Fig. 9. Per Class Error.

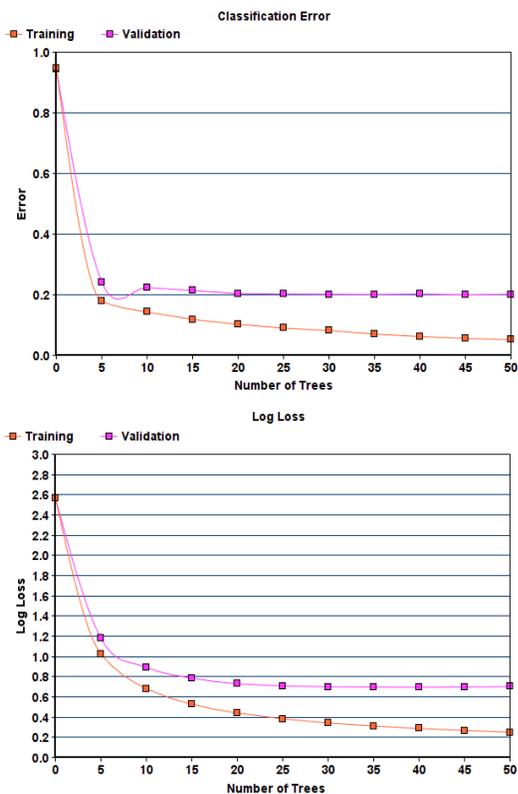


Fig. 10. Classification Error and Log Loss of Question Classifier.

Out of these algorithms XGBoost had the minimum per class error and was chosen as the algorithm of choice for the problem classifier component in the implementation of the system. The classification error and log loss of the algorithm are illustrated in Fig. 10. The figure illustrates the changes in classification error and log loss with respect to the number of trees in the iteration of the algorithm for training as well as validation data.

This problem of imbalanced classes refers to the situation that all the subjects have not been represented equally. More precisely, there is a huge disproportionality between classes, with certain subjects having more than 800 samples and some not even having 80 samples. The problem of imbalanced classes has been greatly explored in excellent reviews in [33], [34]. Class weights were calculated from the dataset distribution frequency and the model was retrained with the best hyperparameters found by the previous search. As a result, the F1 score as well as precision and recall values improved as seen in Table I.

In the quest of proposing a general-purpose approach to assigning user-submitted questions their appropriate subject

labels, without any extra vocabulary information related to the subjects, several grid hyperparameter-searched iterations of Generalized Linear Model, Deep Learning (ANN), Gradient Boosting Model, Extreme Gradient Boosting Model (XGBoost), Distributed Random Forests and Extremely Randomized Trees were trained and evaluated, and it was found that out of these, XGBoost performs the best with a small and imbalanced dataset of very short text documents.

TABLE I. CLASSIFICATION REPORT FOR XGBOOST

Class	After Class Weights			Support
	Precision	Recall	F1-Score	
Artificial Intelligence	0.29	0.14	0.19	14
Computer Networks	0.87	0.83	0.85	168
Digital Logic	0.79	0.86	0.83	143
Algorithms	0.72	0.72	0.72	109
Computer Organization	0.73	0.79	0.76	112
Operating Systems	0.78	0.75	0.76	167
C Programming	0.79	0.71	0.75	105
Data Structures	0.69	0.8	0.74	108
Software Engineering	0.87	0.79	0.83	78
Compiler Design	0.74	0.73	0.73	73
DBMS	0.86	0.87	0.86	166
Theory of Computation	0.89	0.88	0.88	131
Computer Graphics	0.13	0.18	0.15	11

To further improve the classification performance, a general-purpose approach to handle the unbalanced classes was used utilizing class weights. The overall average performance of XGBoost on the validation data has been shown in Table II.

The workflow also same training can be repeated when implementing the system for any other subject domain. Through the ability of problem classification, the system is able to classify new and unseen learner submitted problem to their appropriate topics, and if there aren't any appropriate topics, the system allows the subject experts on the backend to create new more appropriate topics for the problem at hand, resulting in the expansion of the system.

B. Pedagogy Module and Student Module

Based on the topic relationships illustrated in Fig. 3, learner response data was collected from 60 learners assessed over 10 topics and using the algorithm proposed in [32] the precision of the relative prediction of knowledge of other topics when knowledge of one topic is given was calculated as illustrated in Table III. The intensity of the shade in each cell denotes the magnitude of the precision.

TABLE II. PERFORMANCE METRICS OF XGBOOST ON THE CLASSIFICATION OF VALIDATION DATA

	Before Class Weights Adjustment			After Class Weights Adjustment			Support
	Precision	Recall	F1-Score	Precision	Recall	F1-Score	
Macro Average	0.69	0.68	0.68	0.7	0.7	0.7	1385
Weighted Average	0.79	0.8	0.79	0.79	0.79	0.79	1385

V. CONCLUSION

In this paper, an architecture for a domain-independent intelligent tutoring system that features extensibility, personalization and automatic course improvements has been described. It features an interface module that provides separate interfaces for learners and subject experts. The architecture allows expansion of the system without requiring persistent intervention or any knowledge of building ITSs from the subject experts and portable (mobile) interfaces for learners for better learning, motivation and engagement. This provides countless avenues for cost-effective tutoring.

The proposals for future work would be to add support in the architecture that allows the ability of creation of new portable modules that could be built by anyone using a possibly defined format, and added to the system to incorporate various features into any part of the system. An example would be a module that replaces the pedagogy algorithms, or a module that utilizes neural networks for a part of the system, etc. Also proposed is the ability to integrate with other apps, since this architecture has an interface module that works through APIs, the possibility of what can be achieved after integration with other apps and platforms would be exciting to explore.

REFERENCES

- [1] Januszewski and M. Molenda, Educational technology: a definition with commentary. Association for Educational Communications and Technology, 2008.
- [2] V. J. Shute and J. Psotka, "Intelligent Tutoring Systems: Past, Present, and Future.," Armstrong Lab Brooks AFB TX Human Resources Directorate, 1994.
- [3] C. Dede and K. Swigger, "The evolution of instructional design principles for intelligent computer-assisted instruction," *J. Instr. Dev.*, pp. 15–22, 1988.
- [4] J. R. Carbonell, "AI in CAI: An Artificial-Intelligence Approach to Computer-Assisted Instruction," *IEEE Trans. Man-Mach. Syst.*, vol. 11, no. 4, pp. 190–202, Dec. 1970, doi: 10.1109/TMMS.1970.299942.
- [5] J. H. Larkin and R. W. Chabay, "Computer-assisted instruction and intelligent tutoring systems - shared goals and complementary approaches," undefined, 1992.
- [6] M. Elsom-Cook, O. University, and M. K. (GB) C. A. L. R. G. O. University, Intelligent Computer-aided Instruction Research at the Open University. Open University, 1987. [Online]. Available: <https://books.google.co.in/books?id=ptCgHAAACAAJ>.
- [7] J. W. Schofield, R. Eurich-Fulcer, and C. L. Britt, "Teachers, computer tutors, and teaching: The artificially intelligent tutor as an agent for classroom change," *Am. Educ. Res. J.*, vol. 31, no. 3, pp. 579–607, 1994.
- [8] W. L. Miller, R. S. Baker, M. J. Labrum, K. Petsche, Y.-H. Liu, and A. Z. Wagner, "Automated detection of proactive remediation by teachers in Reasoning Mind classrooms," in Proceedings of the Fifth International Conference on Learning Analytics and Knowledge, 2015, pp. 290–294.
- [9] D. Diziol, E. Walker, N. Rummel, and K. R. Koedinger, "Using intelligent tutor technology to implement adaptive support for student collaboration," *Educ. Psychol. Rev.*, vol. 22, no. 1, pp. 89–102, 2010.
- [10] A. C. Graesser, K. VanLehn, C. P. Rosé, P. W. Jordan, and D. Harter, "Intelligent tutoring systems with conversational dialogue," *AI Mag.*, vol. 22, no. 4, pp. 39–39, 2001.
- [11] K. R. Koedinger, A. Corbett, and others, Cognitive tutors: Technology bringing learning sciences to the classroom. na, 2006.
- [12] H. Crompton, "A historical overview of mobile learning: Toward learner-centered education.," *Handb. Mob. Learn.*, no. August 2013, pp. 3–14, 2013.
- [13] G. Trentin and M. Repetto, Using Network and Mobile Technology to Bridge Formal and Informal Learning. Chandos Publishing, 2013. doi: 10.1533/9781780633626.
- [14] N. Bukharaev and A. W. Altaher, "Mobile Learning Education has Become More Accessible," *Am. J. Comput. Sci. Inf. Technol.*, vol. 5, no. 2, Oct. 2017, doi: 10.21767/2349-3917.100005.
- [15] R. Freedman, S. S. Ali, and S. McRoy, "What is an Intelligent Tutoring System?," *Int. J. Artif. Intell. Educ.*, vol. 11, no. 3, pp. 15–16, Sep. 2000, doi: 10.1145/350752.350756.
- [16] R. Nkambou, R. Mizoguchi, and J. Bourdeau, Eds., Advances in Intelligent Tutoring Systems. Berlin Heidelberg: Springer-Verlag, 2010. Accessed: Aug. 08, 2019. [Online]. Available: <https://www.springer.com/gp/book/9783642143625>.
- [17] H. S. Nwana, "Intelligent tutoring systems: an overview," *Artif. Intell. Rev.*, vol. 4, no. 4, pp. 251–277, Dec. 1990, doi: 10.1007/BF00168958.
- [18] S. J. Derry, L. W. Hawkes, and U. Ziegler, "A plan-based opportunistic architecture for intelligent tutoring," *Proc. Intell. Tutoring Syst. ITS-88*, pp. 116–123, 1988.
- [19] J. Siemer and M. C. Angelides, "A comprehensive method for the evaluation of complete intelligent tutoring systems," *Decis. Support Syst.*, vol. 22, no. 1, pp. 85–102, 1998.
- [20] C. Dede, "A review and synthesis of recent research in intelligent computer-assisted instruction," *Int. J. Man-Mach. Stud.*, vol. 24, no. 4, pp. 329–353, 1986.
- [21] E. de Barros Costa and A. Perkusich, "Modeling the cooperative interactions in a teaching/learning situation," in International Conference on Intelligent Tutoring Systems, 1996, pp. 168–176.
- [22] T. E. Turner, M. A. Macasek, G. Nuzzo-Jones, N. T. Heffernan, and K. R. Koedinger, "The Assistent Builder: A Rapid Development Tool for ITS.," in AIED, 2005, pp. 929–931.
- [23] A. C. Graesser et al., "AutoTutor: A tutor with dialogue in natural language," *Behav. Res. Methods Instrum. Comput.*, vol. 36, no. 2, pp. 180–192, 2004.
- [24] C. I. Hausladen, M. H. Schubert, and E. Ash, "Text classification of ideological direction in judicial opinions," *Int. Rev. Law Econ.*, vol. 62, p. 105903, Jun. 2020, doi: 10.1016/J.IRLE.2020.105903.
- [25] J. P. Haumahu, S. D. H. Permana, and Y. Yaddarabullah, "Fake news classification for Indonesian news using Extreme Gradient Boosting (XGBoost)," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 1098, no. 5, p. 052081, Mar. 2021, doi: 10.1088/1757-899X/1098/5/052081.
- [26] A. S. Khan, H. Ahmad, M. Z. Asghar, F. K. Saddozai, A. Arif, and H. A. Khalid, "Personality Classification from Online Text using Machine Learning Approach," *Int. J. Adv. Comput. Sci. Appl.*, vol. 11, no. 3, 2020, doi: 10.14569/IJACSA.2020.0110358.
- [27] T. Joachims, Learning to classify text using support vector machines, vol. 668. Springer Science & Business Media, 2002.
- [28] SebastianiFabrizio, "Machine learning in automated text categorization," *ACM Comput. Surv. CSUR*, vol. 34, no. 1, pp. 1–47, Mar. 2002, doi: 10.1145/505282.505283.
- [29] C. C. Aggarwal and C. Zhai, Mining text data. Springer Science & Business Media, 2012.
- [30] A. Chaturvedi, S. Yadav, M. A. M. H. Ansari, and M. Kanojia, "Comparative Multinomial Text Classification Analysis of Naïve Bayes and XGBoost with SMOTE on Imbalanced Dataset," pp. 339–349, 2021, doi: 10.1007/978-981-16-2543-5_29.
- [31] S. Singh and V. Singh, "Mapping User-submitted Short Text Questions to Subjects of Study: A Multinomial Classification Approach," presented at the 3rd International Conference on Communication and Intelligent Systems (ICCIS 21), National Institute of Technology, Delhi, Dec. 2021.
- [32] S. Singh and V. Singh, "A Graph Based Approach to Learner Profiling in an Intelligent Tutoring System," *Indian J. Comput. Sci. Eng.*, to be published.

- [33] S. Kotsiantis, D. Kanellopoulos, and P. Pintelas, "Handling imbalanced datasets: A review," *GESTS Int. Trans. Comput. Sci. Eng.*, vol. 30, no. 1, pp. 25–36, 2006.
- [34] S. Datta and A. Arputharaj, "An Analysis of Several Machine Learning Algorithms for Imbalanced Classes," in 2018 5th International Conference on Soft Computing Machine Intelligence (ISCMI), 2018, pp. 22–27. doi: 10.1109/ISCMI.2018.8703244.

Secure Routing Protocol for Low Power and Lossy Networks Against Rank Attack: A Systematic Review

Laila Al-Qaisi, Suhaidi Hassan, Nur Haryani Binti Zakaria
InterNetWorks Research Lab, School of Computing
Universiti Utara Malaysia, Kedah Darul Aman, Malaysia

Abstract—The Internet of Things (IoT) is witnessing massive widespread along in almost all aspects of life. IoT is defined as a network of interconnected devices applied in various environments including smart cities, transportation, health, industries, military, and agriculture. Its main purpose is to simplify the exchange and collect data from and to deployment environments. Due to their small size and cost-effectiveness, Wireless Sensor Networks (WSN) form one of the core technologies deployed in IoT. Yet, things interconnected with each other and exchanging data are prone to different kinds of security attacks. As a result, it is possible to compromise data while transmitted from source to destination through nodes. Routing Protocol for Low Power and Lossy Networks (RPL) offers only slight protection against routing attacks, but having a network with limited energy sources, processors, and memory, besides being deployed in unattended nature and hostile environment requires more scalable security measures. This paper focuses on investigating the problem of security provisioning in RPL. As such, a Systematic Literature Review (SLR) of security mechanisms proposed for RPL will be discussed. An extensive search was conducted on various online databases, then findings were filtered by reviewing abstracts, introduction, and conclusion. Finally, a summary of recent research work is presented. This work is important to highlight various aspects of securing RPL and get an initial insight for studying them.

Keywords—Wireless sensor networks; internet of things (IoT); routing security; RPL; objective function

I. INTRODUCTION

Internet of Things (IoT) emergence was led by the assistance of existing wireless communications along with Radio-Frequency Identification (RFID), Wireless Sensor Network (WSN) technologies besides new emerging technologies such as Information-Centric Network (ICN) and Named Data Networks (NDN) [1]. So, data is easily transmitted between various devices and associative things regardless of time and place through network standards and protocols. Every device and thing in IoT is assigned a unique Internet Protocol (IP) address, by which they can sense and collect data from the deployment environment for both processing and decision making. IoT is contributing significantly to various domains like smart cities, building, healthcare, and agriculture and has a vital impact on improving people's daily life [2].

IoT architecture is presented in the literature as mentioned by [3]–[5] consisting of three main layers, namely, perception, network, and application layers. As a hot research topic, many

researchers found the three layers architecture very basic and is suitable for defining the main terminology of IoT and cannot be used for research that digs into further components of IoT. This is when the five layers architecture was introduced as [3] explained, it included processing and business as additional layers. Fig. 1 shows both three- and five-layers architecture.

The network layer is responsible for communication and information exchange employing techniques, standards, and protocols to simplify the task such as Internet Protocol Version 4 (IPv4), Internet Protocol Version 6 (IPv6), Constrained Application Protocol (CoAP), Wireless Personal Area Network (WPAN), IPv6 over Low Power Wireless Personal Area Network (6LoWPAN), User Datagram Protocol (UDP) and Transmission Control Protocol (TCP). Securing data transmitted between the perception layer and the application layer is facilitated by the network layer as well [6].

The Routing Protocol for Low-Power and Lossy Networks (RPL) was developed by the Internet Engineering Task Force (IETF) to fit into Wireless Sensor Networks (WSNs) and the Internet of Things (IoT) domains. As a simple networking protocol, RPL was designed as an interoperable protocol that handles resource-constrained devices connected via multi-hop networks. It enables efficient use of smart devices' energy along with the establishment of flexible topology and routing of data [7].

Nevertheless, the RPL protocol since its inception suffers from a lack of security measures at the network layer as stated by [8]. RPL and its improved versions suffer from a severe performance gap towards network attacks especially ranking attacks [9]. Securing IoT routing should be studied considering WSN features as they are inherited into the IoT environment [10]. Moreover, other metrics in RPL should be taken into consideration such as power consumption as a major challenge facing IoT and controls network lifetime [11].

Cryptography, Intrusion Detection Systems (IDS), Intrusion Prevention Systems (IPS), authentication, trust-based mitigation techniques, and much more, have all been introduced to solve security vulnerabilities in LLNs [12]. In the application, transport, network, and physical levels, IoT devices and traditional PCs share some similar protocols. The biggest impediment to LLN devices implementing existing security methods at IoT interfaces is their limited computational and energy resources [13]. LLN devices produce massive amounts of data, but they lack the resources to store and process it.

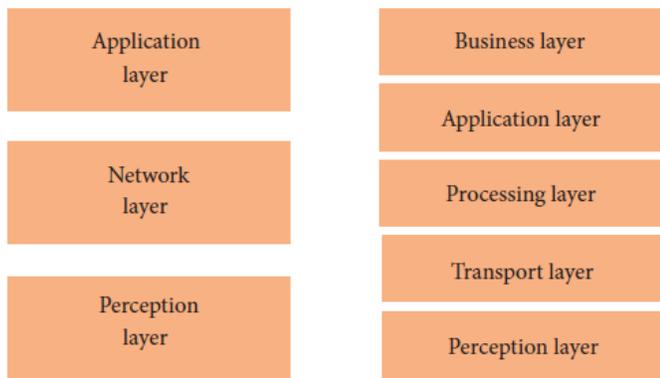


Fig. 1. IoT Architectures [3].

Since in IoT, RPL plays a vital and broad role in service providing, it's a clear target for attackers and a crucial candidate for defense as well. To overcome security issues and challenges in RPL routing, further research is required as per [14], and how intrusions on RPL can be detected is one facet of defense that must now be examined. As a result, this forms a starting point to investigate, propose, and implement mitigation mechanisms for network layer attacks [15].

The main goal of this study is to show the impact that rank attacks (RA) can have on RPL networks. Also, to study and compare the available research that support RPL security and counter the effect of these attacks in terms of the security techniques utilized and their performance. To point up the flaws in the available solutions suggested by existing studies. To suggest some potential methods to address the existing flaws and increase RPL security in IoT networks by limiting the effects of RA. Also, discuss some open challenges in this study area that require more attention.

This paper presents an SLR of security mechanisms proposed for RPL RA specifically being one of the most destructive attacks targeting RPL topology. Starting with RPL in-depth explanation. Followed by a discussion on RPL attacks along with a suitable taxonomy. A focus on rank attacks is presented afterward. Finally, a summary of the selected studies is presented. The remainder of this paper is organized as follows: Section 2 defines and explains preliminaries. Then, Section 3 identifies and explains the following SLR methodology. Section 4 discusses the results found thoroughly. Finally, Section 5 summarizes selected research papers and therefore compares the approaches used by the researchers.

II. PRELIMINARIES

A. Routing Protocol for Low-Power and Lossy Networks (RPL)

Since Low-Power and Lossy Networks (LLN) consist of highly constrained devices in terms of memory, processing capabilities, and energy resources, RPL was designed as an IPV6 distance vector protocol to support communication among LLN devices such IoT. It was mentioned by [16] and [17] that such networks suffer from low data and packet delivery rates along with lossy connection which RPL was designed to be flexible enough for network conditions'

adaptation and provide suitable alternative routes when default ones are not available for any reason at any time.

RPL can be defined as a proactive routing protocol that relies on the distance between nodes and sink node to form a topology. The following explanation of the RPL hierarchy is based on [18]–[21]:

1) *Hierarchy*: Using the distance vector procedure RPL exploits Directed Acyclic Graphs (DAG) mechanism to construct a structure tree or DODAG (Destination Oriented Directed acyclic graph) that controls available nodes' connections with each other. This will enable multi-hop communication via the closest nodes.

RPL methods for establishing connections include point-to-point (P2P), point-to-multipoint (P2MP), and multipoint-to-multipoint (MP2P) communications. While types of nodes for constructing topology are, the source that are responsible nodes for gathering information, leaf nodes that do not perform any task and sink nodes which are the most significant with capabilities of energy and processing to compile whole network information. Hence, two major terms are required here, Control Messages (CM) by which connections are initiated and maintained along with topology formation, and Objective Functions (OF) for routing decision making through the network.

Four types of CM are used to exchange information between nodes in RPL:

- DODAG Information Solicitation (DIS): it is used to request passing the DIO to network neighbors.
- DODAG Information Object (DIO): Stores pertinent information needed to build upward DODAGs route such as RPLInstanceID, configuration parameters, candidate parent information, DODAG maintenance, and more.
- Destination Advertisement Object (DAO): sends information to register every node visited on the downward route.
- Destination Advertisement Object Acknowledgement (DAO-ACK): confirms safe receipt of sent DAO message to the sender node.

2) *Objective function (OF)*: OF was described as the basic element that is handling several vital definitions;(1) computing link cost, (2) parent node selection (when, who, and how many candidates), and (3) computing rank cost, fourth: advertising path cost. There are two defaults OF with RPL, MRHOF (Minimum Rank with Hysteresis Objective Function) and OF0 (Objective Function Zero), and the following are their definitions as per [22]–[24]:

- OF0: This OF adds a specifically predefined value to the previous rank. It takes hop count as a routing metric and selects the best parent node from available candidates based on that. While building the DODAG, nodes should consider hop count to get the shortest path for reaching the grounded root. The rank increases

while going down from root to candidate nodes. However, reliance on node metrics will cause poor link quality. Also, selecting the shortest path in terms of minimum hop count may lead to more retransmissions along with increased packet loss if the path was unreliable. Additionally, this same shortest path may cause more node failure which will definitely decrease network lifetime.

- MRHOF: This OF was designed to overcome the shortcoming of OF0 which depends on a single node metric to compute rank and choose the best parent node. It relies on the expected transmission count (ETX) as a dynamic link metric to stabilize the rank. Still, it chooses the lowest-cost path and avoids network churn overflow using two mechanisms. First, choose a low-rank path, and second hysteresis mechanism ensures changing rank to a lower one only if there exists a rank that is less than the current one. Literature has two main implementations of MRHOF, one that relies on ETX and the other relies on energy.

3) *Routing metrics*: Routing metrics are essential to evaluate path cost and then choose the lowest cost path. There are too many implementations in literature for OF, some take a single metric to calculate rank, while others consider more than one metric. As a matter of fact, metrics can be categorized based on their characteristics into node and link, dynamic and static, quality and quantity routing nodes [25]. Both routing metrics and constraints are used to form a criterion to choose the optimal path. Yet, the main difference between them is that constraint is used to restrict options such as avoiding unreliable links, while metrics define a certain level of reliability to include links that give the optimal path. As a result, both metrics and constraints are used and deployed as per RPL implementation requirements [26]. Moreover, dynamicity is a vital characteristic of metrics, since RPL operating environment is rapidly changing which results in instability of both node and link metrics [27].

The following list summarizes metrics of both link and nodes (refer to Fig. 2):

- Link metrics:

a) *RSSI and LQI*: main radio link estimators are the Received Signal Strength Indicator (RSSI) and the Link Quality Indicator (LQI). The former indicates the level of power received by an antenna that is a high level of RSSI means a stronger radio signal which indicates a closer destination. While the latter measures the quality of the link using a range of values between 0 to 7.

b) *ETX*: Expected Transmission Count indicates the reliability of the network and gives the required number of transmissions for receiving acknowledgment from the destination.

- Node metrics:

a) *Energy*: represents the energy consumed by nodes through network operations.

b) *Hop count*: it is a measure of path link that is used extensively in wireless networks and the main drawback is to get the shortest path with the lowest hop count regardless of link quality.

c) *End-to-end delay*: a vital metric for building route in RPL and it indicates the needed time to deliver packets to the sink from sender nodes.

B. RPL Attacks Classification

RPL is vulnerable to various kinds of attacks and does not have a solid security measure that can prevent such attacks [28]. There are several taxonomies proposed for attacks targeting RPL in different studies, such as Almusaylim et al. [17] in which three main types of attacks were explained namely; against resources that consume nodes resources, topology in which try to cause damage in the construction process and traffic which aim at capturing as much traffic as possible. Also, Avila et al. [10] categorized attacks into passive and active attacks, where passive attacks aim to gather information after accessing the system and comprise confidentiality, and active ones sabotage the system by data alteration, disabling nodes, or giving access to unauthorized users. An interesting categorization was presented by Raouf et al. [29], in which attacks were classified based on their origin into RPL Specific and WSN inherited as Fig. 3 shows.

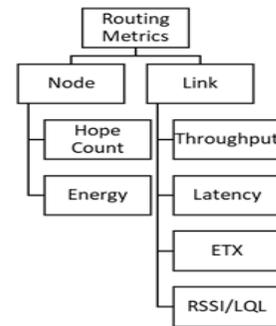


Fig. 2. Routing Metrics.

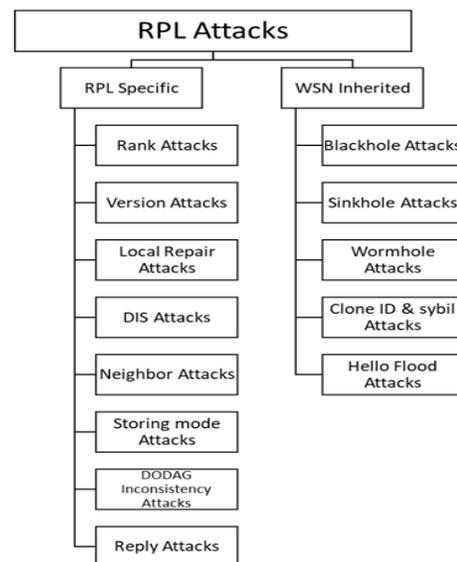


Fig. 3. RPL Attacks.

III. METHODOLOGY

This study employed the systematic literature review guidelines and standards proposed by Kitchenham [30]. This consists of a set of well-defined stages conducted in line with a predefined protocol. SLR consists of three phases: planning, conducting, and reporting the reviews according to Shaffril et al. [31]. These phases consist of the following processes: (1) identifying RQs; (2) developing a review protocol; (3) determining both exclusion and inclusion criteria; (4) selecting search strategy and study process; (5) quality assessment (QA); and (6) extracting and synthesizing data.

A. Identifying Research Questions (RQs) Text

To achieve the main objectives of this study, primary studies should be assessed and reviewed thoroughly. As a result, the following research questions are proposed based on Population, Intervention, Comparison, Outcomes, and Context (PICOC) as per [30]:

- RQ1: What is the impact of the rank attack and to which extent do they damage the network?
- RQ2: What are the proposed approaches that monitor the network to handle attacks targeting RPL?
- RQ3: What are the technical performance metrics of the research in this field?
- RQ4: What are the advantages and disadvantages of each proposed approach?

B. Developing a Review Protocol

A vital step that makes SLR different from traditional methods of reviewing the literature. Because it decreases study bias as discussed by Shaffril et al. [32]. The review protocol categorizes review background, search strategy, development of RQs, extraction of data, criteria for study selection, and data synthesis.

C. Search Strategy

The search strategy started with choosing E-digital libraries and online databases as the following list shows, taking into consideration selecting only high-impact-factor publications:

- IEEE Explore
- ACM Digital Library
- Science Direct
- Scopus
- Wiley Online Library

Afterward, the search string is required to conduct an in-depth search through selected E-digital libraries. The following steps were applied to define the used search string as per [30]:

- Define major keywords depending on identified research questions.
- Consider linguistic synonyms, alternatives, and interchangeable terms for each keyword.
- Use conjunction operators (AND, OR) when needed to produce the full search string.

As a result, keywords included for the search were “IoT” OR “Internet of Things” AND “RPL” OR “Routing Protocol for Low-Power and Lossy Networks” AND “rank attack detection” OR “mitigation”. All available papers relating to specified keywords 2022 were collected from digital libraries.

Afterward, a manual search was applied to the results of the automatic search by filtering each paper's title, abstract and content. This is to ensure that the selected paper supports answering the defined QAs and Fig. 4 illustrates the overall search phases.

D. Inclusion and Exclusion Criteria

Search results are filtered in terms of the following inclusion and exclusion criteria:

- Inclusion Criteria:

- a) Written in English language.
- b) The study's domain is RPL and responds to previously stated RQs.
- c) Published in journal or conference.
- d) Published date: 2017-2022.

- Exclusion Criteria:

- a) Duplicates.
- b) Unavailable full text.
- c) Do not meet the inclusion criteria.

Afterward, a manual filtration process was conducted by reviewing the title, abstract, and conclusion to get papers that meet the set criteria of found papers. This eliminated the number of found papers from 1061 to 9 only, given that only papers published between 2017 to 2022 and studied RA in RPL only.

E. Applying Quality Assessment (QA)

The related studies' quality was assessed using QA as recommended by Kitchenham [30]. All found studies were assessed concerning every single research question. QA criteria used for the assessment process were as follows:

- QA1: Is the topic addressed in the paper related to securing RPL?
- QA2: Is there any mechanism proposed to detect rank attack detection in RPL?
- QA3: Is there a sufficient explanation of the background in which the study was performed?
- QA4: Is there a clear declaration concerning methods used to validate the applied mechanism?

The reliability of articles and studies found was tested through the four QA criteria and has three categories low, medium, and high as by Shaffril et al. [31] and [32]. Each QA had a score of 2 points and each paper that meets the defined QA earns a score of 2, 1 is earned when the paper partially meets the QA criteria and 0 when it does not satisfy the QA criteria at all. Papers scored more than 5 are discussed in the next section and are categorized based on the technique used and Table II summarizes the findings sorted by year of publication.

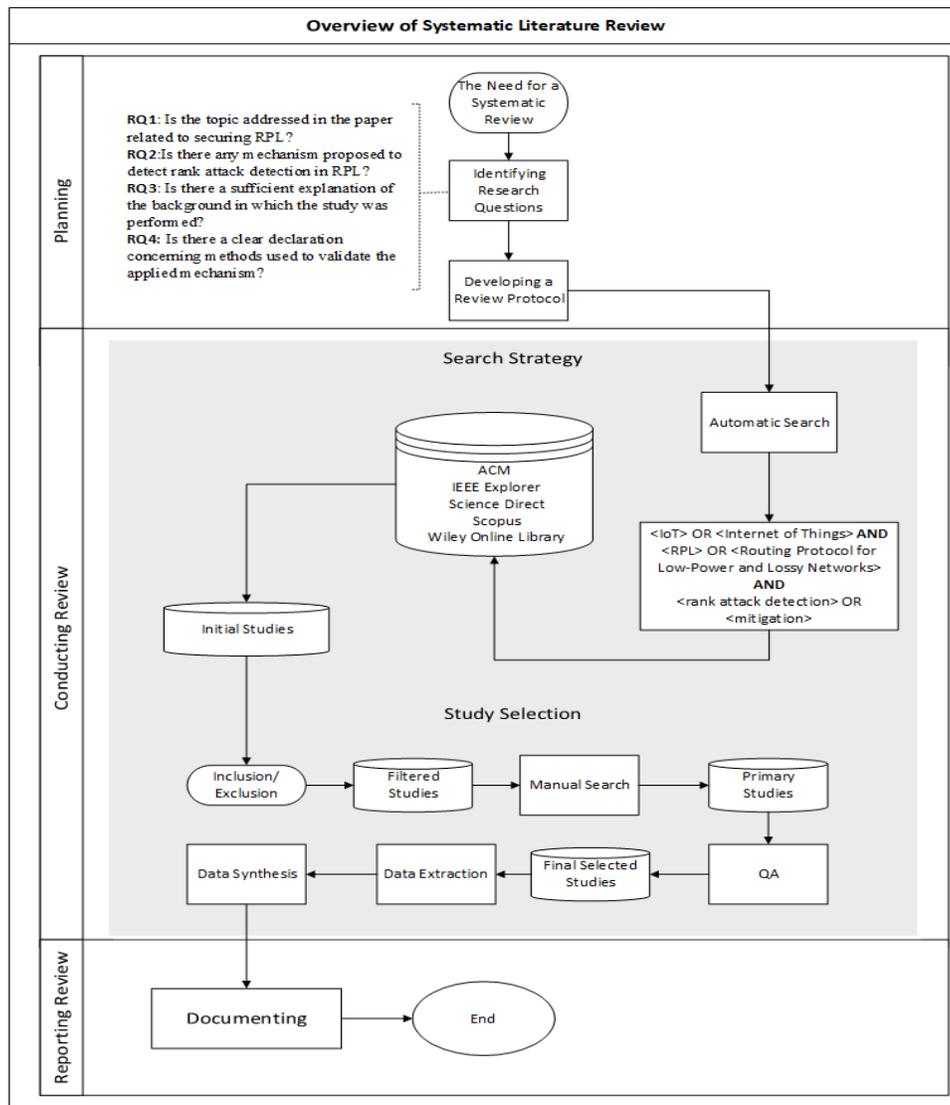


Fig. 4. Systematic Literature Review (SLR).

F. Data Extraction and Synthesis

For accurate data extraction and synthesis, a form was developed to conduct this step. Details of each study related to the reference, year of publication, methodology, and comments were extracted. A tabular form was used to register this information about each study. Table I illustrates the details registered for each paper.

TABLE I. TABLE TYPE STYLES

Extracted Data	Description
Study ID	paper DOI
Year	Publication Date
Type	Journal or conference
Methodology	e.g., trust, cryptography, IDS
Performance Measure	e.g., ACC, PR, RE

IV. RESULTS

This section discusses rank attacks against RPL and analyzes the application of detection and mitigation techniques towards it. The methods analyzed herein are ones that were proposed to secure RPL against RA. The goal is to present their performance in terms of the chosen performance metrics which will be discussed here as well.

A. Rank Attack (RA) (RQ1)

This attack aims at attracting network traffic to a specified node. Ul Hassan et al. [34] defined RA as an attack that occurs when the malicious node sends information of a lower range, to be closer than others to the root. This scenario will have a consequence that makes malicious nodes able to capture as much traffic as possible. Hashemi and Aliee [35] mentioned that RA is considered the most destructive attack among other types, this is because it intentionally aims at downgrading the network performance by tampering with the rank. By which a rank is decreased to make the malicious node closer to the

chosen parent, so a massive amount of passing packets through it may be manipulated.

RA workflow starts when a malicious node sends a fake rank through an RPL control message or advertises a fake route across the root node to mislead close nodes to make them transmit packets through it [36]. In other words, RA exposes ranks of child nodes in the RPL network topology, then modifies the way of processing DIO messages by neighboring nodes. The worst part will occur when a malicious node with a fake rank is chosen as the preferred parent node while operating, which will result in creating more traffic for data packets to go through the malicious node as un-optimized routing occurs due to network topology OF is not completely achieved as discussed [37].

Mishra and Pandya [38] added another scenario for rank attacks by which an attacker node advertises a better routing metric to other neighboring nodes although it's fake, it misleads network flow to be passing through it. Besides, this may lead to significant increasing latency and decreasing throughput in the network. Fig. 5 illustrates an example of RA.

RA may affect the network and causes several issues as discussed by Nandhini and Mehtre [39]: first, form an unoptimized route. The second is unrecognized loop formation. Third, RPL network topology never uses the optimized route. Fourth, the decreased packet delivery ratio affects the delay increase. Fifth, network topology changes rapidly causing DIO messages number to increase. Some network restricted resource properties would be affected such as energy consumption, throughput, latency, and data rate.

As a result, RPL security forms a major concern that should be considered and further investigated, especially when RA is the topic. This is because routed data shouldn't be accessed by a third party or attacker.

B. RPL Rank Attack (RA) Countermeasures (RQ2)

Many papers categorized countermeasures deployed to secure RPL against attacks, Raoof et al. [29] classified detection and mitigation mechanisms into Acknowledgment-based which depends on sending and receiving acknowledgment messages to prevent any suspicious alteration, and Trust-based depending on the node to monitor neighboring nodes by rating them and consider a ratio to accept, Location-based considering physical location of nodes and Statistical/Mathematical-based by which a mathematical calculation is considered to detect attacks.

Further classification is presented by Verma and Ranga [37] added to the above mentioned, Intrusion Detection Systems (IDS) that consists of signature-based IDS, anomaly-based IDS, and specification-based IDS. It is defined by [40] as, a complete system that may be deployed either in a stand-alone computer system or a network. Its main role is to monitor activities and analyze them to specify any incident which targets security policies integrity, availability, or confidentiality and report it as unauthorized or malicious activity.

Muzammal et al. [41] also mentioned IDS as a significant method that is used to mitigate attacks of RPL in addition to all previously stated ones. Besides many alterations to OF by combining various previously explained link and nodes metrics with adopting additional methods such as fuzzy logic.

Moreover, Tasneem and Wahid [42] classified proposed defense methods for RPL into reactive approaches that include cryptography-based, trust-based, and threshold-based methods, and proactive approaches which consist of time-based and energy-based methods.

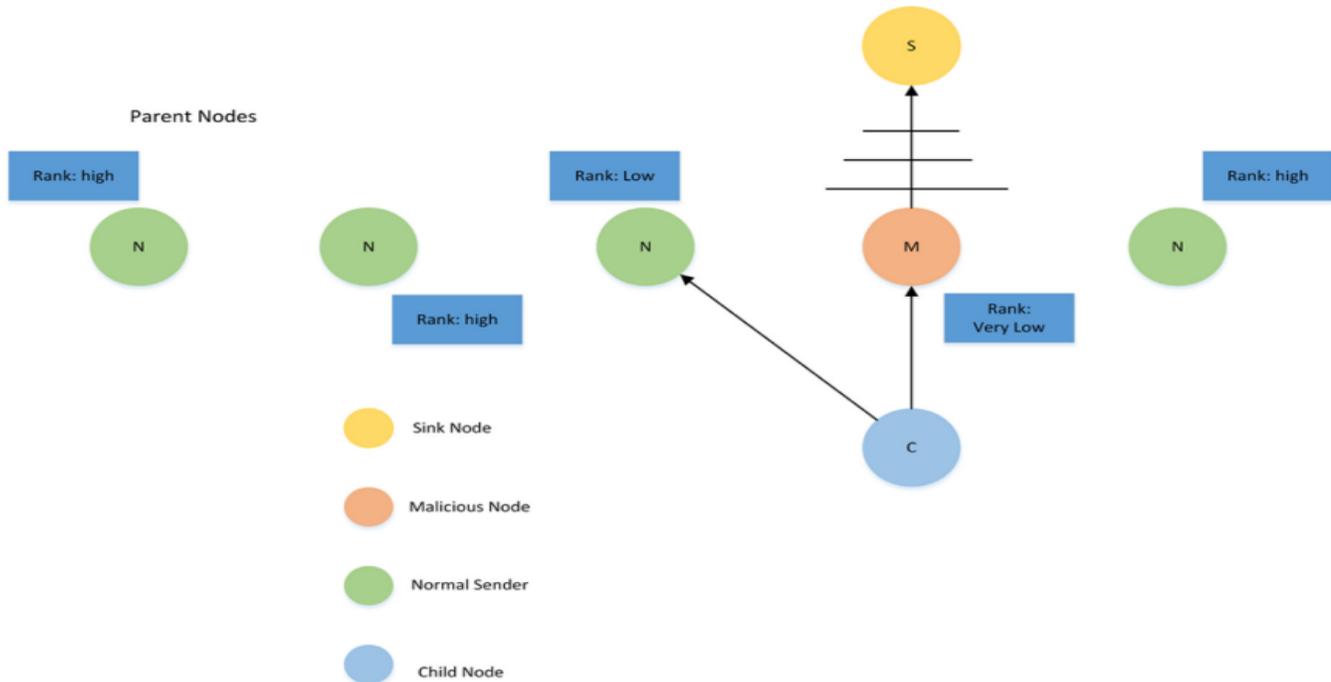


Fig. 5. RA Example, [33].

Finally, countermeasures proposed for RPL against RA were classified by Almusaylim et al. [17] into modification techniques by which some alterations may be applied to a certain component of RPL such as DODAG, OF, or ranking policies and IDS.

C. Performance Metrics (RQ3)

Various metrics were used for measuring the performance of the proposed methods. Yet, many studies like [7], [9], and [14] mentioned using node and link metrics discussed previously such as power consumption, ETX, and PDR. In addition, accuracy metrics including True Positive Rate (TPR), False Positive Rate (FPR), and Detection Rate (DR) were mentioned to be used as well, and below are their formulas as per [43]–[47]:

- Detection Rate (DR): Refers to the ability of the model to rank patterns, and its ability to select a threshold in the ranking used to classify patterns as normal if above the threshold and abnormal if below. It is calculated using Equation 1 below:

$$DR = \frac{TPR}{TPR+FNR} * 100\% \quad (1)$$

where All = TPR + TNR + FPR + FNR

- TPR: Also called sensitivity and it measures the truly predicted positive and were correctly identified. It is calculated as in Equation 3:

$$TPR = \frac{TPR}{TPR+FNR} * 100\% \quad (2)$$

Where FNR is calculated as follows:

$$FNR = \frac{TPR+TNR}{TPR+FPR} * 100\% \quad (3)$$

- FPR: Refers to the probability of a False Alarm. That is, the percentage of actual abnormal flows predicted as normal flows and it is calculated as in Equation 4:

$$FPR = \frac{FPR}{FPR+TNR} * 100\% \quad (4)$$

Where TNR is calculated as follows:

$$TNR = \frac{TNR}{FPR+TNR} * 100\% \quad (5)$$

D. Summary of Shortlisted Studies (RQ4)

This section discusses thoroughly found nine studies that proposed techniques to detect and mitigate RA targeting RPL and strictly meet the criteria defined in the SLR methodology section along with a summary presented in Table II.

A Secure RPL Routing Protocol (SRPL-RP) for rank and version number attacks was proposed by Almusaylim et al. [48] in which a timestamp is added to ensure the legitimacy of sending nodes. A monitoring table is included through the process of constructing DODAG which collects all information about existing nodes. A blacklist and alert tables were added to simplify the procedures of mitigating and isolating both types of studied attacks. Several conditions were added to control the current rank of nodes and parent nodes to maintain a safe network. Simulations were conducted using the Cooja simulator and results showed that the proposed SRPL-RP had a

higher PDR and a lower control message value compared to methods previously proposed in literature along with 95% accuracy in all kinds of tested network topologies.

Shafique et al. [49] proposed a novel sink-based IDS (SBIDS) by which a timespan is added to the DAO message for ensuring its freshness. Then several detection steps are followed to detect any violence in rank, especially a rule that compares node current rank (NCR) to node parent rank NPR. Simulations were conducted using the Cooja simulator and performance metrics were percentage of accuracy, TP, TN, FP, FN, and confidence interval (CI) under mobility conditions. Results showed that SBIDS had 100% detection accuracy under normal circumstances, yet it decreased with the number of nodes with mobility increased.

TABLE II. FINAL SELECTED PAPERS FOR SECURING RPL AGAINST RA

Ref	Paper Information			
	Year	Type	Methodology	Performance Measure
Almusaylim et al. [48]	2020	Journal	SRPL-RP based on rank strategy	PDR, Acc
Shafique et al. [49]	2018	Journal	Sink-based IDS (SBIDS)	Acc, TPR, TNR, FPR, FNR, and confidence interval (CI)
Boudouaia et al. [50]	2021	Journal	Rank property + DIO messages with 2 rank thresholds	DR, average network hops, and global energy consumption.
Nair and BJ [51]	2021	Conference	SCF and Dijkstra's algorithm	Throughput and PDR
Karmakar, Sengupta and Bit [52]	2021	Conference	DODAG modification with adding Authentication Code (HMAC-LOCHA)	DR, FPR, FNR, and energy consumption
Zarzoor [53]	2021	Journal	Layering mechanism	Latency, energy consumption, and DR
Stephen and Arockiam [54]	2018	Conference	Node energy based E2V architecture with IDS	Network convergence delay, energy consumption, and attacker identification delay
Seth et al. [55]	2020	Conference	Round trip time (RTT) based detection and isolation mechanism	Acc
Althubaity, Gong, and Raymond [56]	2020	Conference	Specification-based IDS (FORCE)	DR and overheads incurred on the nodes' resources

Boudouaia et al. [50] proposed a security scheme that uses a rank property to choose a preferred parent in RPL topology, so any malicious behavior in terms of rank may be detected. Afterward, when the DIO message arrives two values will be calculated to indicate the minimum rank threshold and maximum rank threshold depending on the neighboring rank. As a result, nodes that do not match threshold criteria are blacklisted and the selection process will be held upon legitimate nodes only. Experiments were done using the cooja simulator, 4 scenarios, and performance evaluations were conducted in terms of successful detection rate, the average network hops, and the global energy consumption.

Another solution was proposed by Nair and BJ [51] in which both spatial correlation function (SCF) and Dijkstra's algorithm were applied to select the preferred parent nodes using proactive routing in terms of throughput and energy as selection parameters. For experiments, the NS-2 simulator was used, and performance evaluation was based on throughput and PDR.

An interesting study by Karmakar, Sengupta, and Bit [52] combined several methods to secure RPL against RA. First, the algorithm forming DODAG was modified to be able to detect RA during building and maintaining the topology. Second, two modules were added, distributed at all nodes, and centralized at the sink node. Third, the DAO control message was modified to lower overhead levels and a lightweight Message Authentication Code (HMAC-LOCHA) was used to verify exchanged message's integrity and authenticity. Cooja simulator was used to conduct experiments and multiple test case scenarios were applied. detection accuracy, false positive/negative rate, and energy consumption.

Zaroor [53] proposed a security mechanism that relies on the layering principle. It consists of three main phases: first, nodes are categorized into layers. Second, calculate the trust value for the path. Third, detect and mitigate the RA. For implementation Cooja simulator was used and performance evaluation was conducted based on latency, nodes' energy consumption, and accuracy of malicious node detection.

A further three-phase mechanism called E2V was proposed by Stephen and Arockiam [54] which starts with rank calculation, substantiation, and elimination. Where the malicious node is detected at the substantiation phase by the defined IDS. Then, in the elimination phase, malicious nodes will be eliminated by either local repair or global repair. The Cooja simulator is used for implementation purposes and evaluation in terms of network parameters such as network convergence delay, energy consumption, and attacker identification delay.

Seth et al. [55] used round trip time (RTT) to detect verify and isolate malicious nodes from the network in RPL. Cooja simulator was used for implementation and performance was evaluated in terms of accuracy where the proposed scheme was found to be better than previous ones.

Althubaity, Gong, and Raymond [56] proposed a fully distributed specification-based IDS (FORCE). The type of node forms a significant issue for FORCE, yet it was designed so that every single node can analyze and receive control

messages and in case of any attack detection an alert will be generated directly. Evaluation metrics used were detection rates and overheads incurred on the nodes' resources and experiments were conducted using the Cooja simulator.

V. DISCUSSION

The main goal of this part is to understand the obstacles and current research for detecting RA in RPL routing protocols, as well as several flaws that require more research. RPL routing protocols provide for more efficient use of smart devices, resources, and data routing. Because of the characteristics that distinguish this network from others, developing secure routing algorithms for IoT networks is a difficult task. Secure routing techniques for IoT devices have received a lot of attention in recent years. However, they all rely on traditional cryptographic operations, which deplete device resources and have a significant impact on the performance of limited IoT devices. They are vulnerable to a wide range of security threats. The absence of infrastructure, inconsistent links, resource limits, poor physical security, and changing topology of PRLs make them vulnerable to attacks and difficult to defend against.

A. Limitations

Based on reviewed studies it was found that current security features of RPL may be defined but not actually used either in real applications or in research as they are marked as optional features. This puts security as a significant concern of RPL especially since it's being deployed and used widely in IoT environments which are witnessing massive growth globally.

As RPL is vulnerable to several attacks, RA is one of the major attacks that were found to compromise RPL, yet a lower amount of research conducted to specifically target it. Also, these studies had several shortcomings which should be addressed to overcome their consequences.

As a result, it was found from this review: that first, most studies considered either selection or mitigation, but only a few of them investigated both schemes. Second, mainly one type of network topology was selected to test and measure the performance of the proposed scheme. Third, most research studies tend to evaluate their proposed schemes by taking small IoT Networks (<100 nodes) which are considered impractical because the impact of network size on both attacks and security mechanisms remains unknown. Fourth, many schemes encountered an increased number of control messages for acknowledgment purposes which may cause both complexities and increased overhead and are considered inefficient.

B. Comparison

Based on provided review and summery in Table II, it can be concluded that most chosen metrics for performance evaluation were DR and energy consumption as in [50], [52], [53] and [56]. As DR indicated to which extent the proposed mechanism was able to detect threats and energy consumption represented a measure of keeping devices resources available. IDS was chosen as a detection solution in three papers [49], [54] and [56], while the rest choose to modify the main protocol policies and add certain solutions to improve its

security measures. None of found studies tented to combine IDS with protocol policy improvements. Also, none of them included integration with other recently hot fields such as fuzzy logic as a solution.

Studies discussed securing RPL against RA were 5 conference papers to 4 journal papers within period 2017 to 2022, which means this kind of attacks require more powerful solutions are to be proposed in order to provide efficient solution.

Finally, experiments of all founded papers showed that the Cooja simulator usage is dominant in RPL studies where all of them implemented proposed solutions using it.

VI. CONCLUSION

This paper studied applied methods for RA detection in RPL thoroughly to address limitations in this field. An SLR was conducted to determine the required studies to be conducted for improving security measures deployed in this regard. Definitions of required terms starting from IoT, RPL architecture, and security attacks, to detection and mitigation techniques, are presented to help researchers have a brief explanation of them. Also, a summary of recent studies is presented. It was found that many of the currently applied mechanisms in literature have weak points, cryptographic-based methods may provide security, but it definitely consumes nodes' restricted resources. While trust-based may solve the resource restrictions, it may cause other issues regarding network performance such as latency. IDS, it's considered the most effective solution among all proposed ones, but it requires collaboration, and many aspects in this regard should be taken into consideration such as placement. Finally, a hybrid IDS is highly recommended as a solution for securing RPL as it is used by IoT and keeping it safe will definitely be reflected in the overall IoT environment.

VII. FUTURE WORK

Future research aims at extending this review to examine and build better detection and mitigation measures for RPL. This will primarily be addressing RPL rank vulnerabilities.

ACKNOWLEDGMENT

This research was supported by the Ministry of Higher Education (MoHE) of Malaysia through The Fundamental Research Grant Scheme (FRGS/1/2020/ICT07/UUM/01/1).

REFERENCES

- [1] A. Abrar, A. S. B. C. M. Arif, and K. B. M. Zaini, "Producer Mobility Support in Information-Centric Networks: Research Background and Open Issues," *Int. J. Commun. Networks Distrib. Syst.*, vol. 28, no. 1, p. 1, 2022, doi: 10.1504/ijcnds.2022.10044469.
- [2] J. Asharf, N. Moustafa, H. Khurshid, E. Debie, W. Haider, and A. Wahab, "A review of intrusion detection systems using machine and deep learning in internet of things: Challenges, solutions and future directions," *Electron.*, vol. 9, no. 7, 2020, doi: 10.3390/electronics9071177.
- [3] P. Sethi and S. R. Sarangi, "Internet Of Things: Architecture, Issues and Applications," *Int. J. Eng. Res. Appl.*, vol. 07, no. 06, pp. 85–88, 2017, doi: 10.9790/9622-0706048588.
- [4] R. Mondal and T. Zulfi, "Internet of Things and Wireless Sensor Network for Smart Cities," *Int. J. Comput. Sci. Issues*, vol. 14, no. 5, pp. 50–55, 2017, doi: 10.20943/01201705.5055.
- [5] H. Babbar and S. Rani, "Software-defined networking framework securing internet of things," in *Integration of WSN and IoT for Smart Cities*, Springer, 2020, pp. 1–14.
- [6] D. B. D. and F. Al-Turjman, "A hybrid secure routing and monitoring mechanism in IoT-based wireless sensor networks," *Ad Hoc Networks*, vol. 97, p. 102022, Feb. 2020, doi: 10.1016/j.adhoc.2019.102022.
- [7] M. Zaminkar and R. Fotuhi, "SoS-RPL: Securing Internet of Things Against Sinkhole Attack Using RPL Protocol-Based Node Rating and Ranking Mechanism," *Wirel. Pers. Commun.*, vol. 114, no. 2, pp. 1287–1312, Sep. 2020, doi: 10.1007/s11277-020-07421-z.
- [8] M. Pishdar, Y. Seifi, M. Nasiri, and M. Bag-Mohammadi, "PCC-RPL: An efficient trust-based security extension for RPL," *Inf. Secur. J. A Glob. Perspect.*, vol. 31, no. 2, pp. 168–178, Mar. 2022, doi: 10.1080/19393555.2021.1887413.
- [9] S. Y. Hashemi and F. Shams Aliee, "Fuzzy, Dynamic and Trust Based Routing Protocol for IoT," *J. Netw. Syst. Manag.*, vol. 28, no. 4, pp. 1248–1278, Oct. 2020, doi: 10.1007/s10922-020-09535-y.
- [10] K. Avila, D. Jabba, and J. Gomez, "Security Aspects for Rpl-Based Protocols: A Systematic Review in IoT," *Appl. Sci.*, vol. 10, no. 18, p. 6472, 2020, doi: 10.3390/app10186472.
- [11] G. Soni and R. Sudhakar, "A L-IDS against Dropping Attack to Secure and Improve RPL Performance in WSN Aided IoT," in *2020 7th International Conference on Signal Processing and Integrated Networks (SPIN)*, Feb. 2020, pp. 377–383, doi: 10.1109/SPIN48934.2020.9071118.
- [12] T. Park, N. Abuzainab, and W. Saad, "Learning How to Communicate in the Internet of Things: Finite Resources and Heterogeneity," *IEEE Access*, vol. 4, pp. 7063–7073, 2016, doi: 10.1109/ACCESS.2016.2615643.
- [13] D. Midi, A. Rullo, A. Mudgerikar, and E. Bertino, "Kalis — A System for Knowledge-Driven Adaptable Intrusion Detection for the Internet of Things," in *2017 IEEE 37th International Conference on Distributed Computing Systems (ICDCS)*, Jun. 2017, pp. 656–666, doi: 10.1109/ICDCS.2017.104.
- [14] N. Djedjig, D. Tandjaoui, F. Medjek, and I. Romdhani, "Trust-aware and cooperative routing protocol for IoT security," *J. Inf. Secur. Appl.*, vol. 52, p. 102467, Jun. 2020, doi: 10.1016/j.jisa.2020.102467.
- [15] E. Fernandes, A. Rahmati, K. Eykholt, and A. Prakash, "Internet of Things Security Research: A Rehash of Old Ideas or New Intellectual Challenges?," *IEEE Secur. Priv.*, vol. 15, no. 4, pp. 79–84, 2017, doi: 10.1109/MSP.2017.3151346.
- [16] D. B. Gothawal and S. V. Nagaraj, "Intrusion Detection for Enhancing RPL Security," *Procedia Comput. Sci.*, vol. 165, pp. 565–572, 2019, doi: 10.1016/j.procs.2020.01.051.
- [17] Z. A. Almusaylim, A. Alhumam, and N. Z. Jhanjhi, "Proposing a Secure RPL based Internet of Things Routing Protocol: A Review," *Ad Hoc Networks*, vol. 101, p. 102096, Apr. 2020, doi: 10.1016/j.adhoc.2020.102096.
- [18] A. Arena, P. Perazzo, C. Vallati, G. Dimi, and G. Anastasi, "Evaluating and improving the scalability of RPL security in the Internet of Things," *Comput. Commun.*, vol. 151, pp. 119–132, Feb. 2020, doi: 10.1016/j.comcom.2019.12.062.
- [19] M. A. Boudouaia, A. Ali-Pacha, A. Abouaissa, and P. Lorenz, "Security Against Rank Attack in RPL Protocol," *IEEE Netw.*, vol. 34, no. 4, pp. 133–139, Jul. 2020, doi: 10.1109/MNET.011.1900651.
- [20] A. M. Pasikhani, J. A. Clark, P. Gope, and A. Alshahrani, "Intrusion Detection Systems in RPL-Based 6LoWPAN: A Systematic Literature Review," *IEEE Sens. J.*, vol. 21, no. 11, pp. 12940–12968, 2021, doi: 10.1109/JSEN.2021.3068240.
- [21] A. K. Rana and S. Sharma, "Contiki Cooja Security Solution (CCSS) with IPv6 Routing Protocol for Low-Power and Lossy Networks (RPL) in Internet of Things Applications," 2021, pp. 251–259.
- [22] H. Lamaazi and N. Benamar, "A comprehensive survey on enhancements and limitations of the RPL protocol: A focus on the objective function," *Ad Hoc Networks*, vol. 96, p. 102001, Jan. 2020, doi: 10.1016/j.adhoc.2019.102001.
- [23] A. Paul and A. S. Pillai, "A Review on RPL Objective Function Improvements for IoT Applications," *ACCESS 2021 - Proc. 2021 2nd*

- Int. Conf. Adv. Comput. Commun. Embed. Secur. Syst., no. September, pp. 80–85, 2021, doi: 10.1109/ACCESS51619.2021.9563294.
- [24] S. M. M. and D. P. I. Basarkod, "A Comprehensive Survey on RPL: Evolution and Challenges," *SSRN Electron. J.*, 2019, doi: 10.2139/ssrn.3510063.
- [25] G. Violettas, G. Simoglou, S. Petridou, and L. Mamas, "A Softwarized Intrusion Detection System for the RPL-based Internet of Things networks," *Futur. Gener. Comput. Syst.*, vol. 125, pp. 698–714, Dec. 2021, doi: 10.1016/j.future.2021.07.013.
- [26] H. Lamaazi and N. Benamar, "RPL enhancement using a new objective function based on combined metrics," in *2017 13th International Wireless Communications and Mobile Computing Conference (IWCMC)*, Jun. 2017, pp. 1459–1464, doi: 10.1109/IWCMC.2017.7986499.
- [27] S. Sennan and S. Palanisamy, "Composite Metric Based Energy Efficient Routing Protocol for Internet of Things," *Int. J. Intell. Eng. Syst.*, vol. 10, no. 5, pp. 278–286, Oct. 2017, doi: 10.22266/ijies2017.1031.30.
- [28] J. Karlsson, L. S. Dooley, and G. Pulkkis, "Secure Routing for MANET Connected Internet of Things Systems," in *2018 IEEE 6th International Conference on Future Internet of Things and Cloud (FiCloud)*, Aug. 2018, pp. 114–119, doi: 10.1109/FiCloud.2018.00024.
- [29] A. Raoof, A. Matrawy, and C. H. Lung, "Routing Attacks and Mitigation Methods for RPL-Based Internet of Things," *IEEE Commun. Surv. Tutorials*, vol. 21, no. 2, pp. 1582–1606, 2019, doi: 10.1109/COMST.2018.2885894.
- [30] B. Kitchenham, "Procedures for performing systematic reviews," Jun. 2004, [Online]. Available: http://www.elizabete.com.br/rs/Tutorial_IHC_2012_files/Conceitos_RevisaoSistematica_kitchenham_2004.pdf.
- [31] H. A. Mohamed Shaffril, S. F. Samsuddin, and A. Abu Samah, "The ABC of systematic literature review: the basic methodological guidance for beginners," *Qual. Quant.*, vol. 55, no. 4, pp. 1319–1346, 2021, doi: 10.1007/s11135-020-01059-6.
- [32] H. A. M. Shaffril, A. A. Samah, and S. F. Samsuddin, "Guidelines for developing a systematic literature review for studies related to climate change adaptation," *Environ. Sci. Pollut. Res.*, vol. 28, no. 18, pp. 22265–22277, May 2021, doi: 10.1007/s11356-021-13178-0.
- [33] M. Karthik, VK Pushpalatha, "Addressing Attacks and Security Mechanism in the RPL based IOT," *Int. J. Comput. Sci. Eng. Commun.*, vol. 5, no. 5, pp. 1715–1721, 2017.
- [34] T. ul Hassan, M. Asim, T. Baker, J. Hassan, and N. Tariq, "CTrust - RPL: A control layer - based trust mechanism for supporting secure routing in routing protocol for low power and lossy networks - based Internet of Things applications," *Trans. Emerg. Telecommun. Technol.*, vol. 32, no. 3, Mar. 2021, doi: 10.1002/ett.4224.
- [35] S. Y. Hashemi and F. Shams Aliche, "Dynamic and comprehensive trust model for IoT and its integration into RPL," *J. Supercomput.*, vol. 75, no. 7, pp. 3555–3584, 2019, doi: 10.1007/s11227-018-2700-3.
- [36] A. Rehman, M. M. Khan, M. A. Lodhi, and F. B. Hussain, "Rank attack using objective function in RPL for low power and lossy networks," in *2016 International Conference on Industrial Informatics and Computer Systems (CIICS)*, Mar. 2016, pp. 1–5, doi: 10.1109/ICCSIL.2016.7462418.
- [37] A. Verma and V. Ranga, "Security of RPL Based 6LoWPAN Networks in the Internet of Things: A Review," *IEEE Sens. J.*, vol. 20, no. 11, pp. 5666–5690, 2020, doi: 10.1109/JSEN.2020.2973677.
- [38] N. Mishra and S. Pandya, "Internet of Things Applications, Security Challenges, Attacks, Intrusion Detection, and Future Visions: A Systematic Review," *IEEE Access*, vol. 9, pp. 59353–59377, 2021, doi: 10.1109/ACCESS.2021.3073408.
- [39] P. S. Nandhini and B. M. Mehtre, "Intrusion Detection System Based RPL Attack Detection Techniques and Countermeasures in IoT: A Comparison," *Proc. 4th Int. Conf. Commun. Electron. Syst. ICCES 2019*, no. Icces, pp. 666–672, 2019, doi: 10.1109/ICCES45898.2019.9002088.
- [40] B. B. Zarpelão, R. S. Miani, C. T. Kawakani, and S. C. de Alvarenga, "A survey of intrusion detection in Internet of Things," *J. Netw. Comput. Appl.*, vol. 84, pp. 25–37, 2017, doi: 10.1016/j.jnca.2017.02.009.
- [41] S. M. Muzammal, R. K. Murugesan, and N. Z. Jhanjhi, "A Comprehensive Review on Secure Routing in Internet of Things: Mitigation Methods and Trust-Based Approaches," *IEEE Internet Things J.*, vol. 8, no. 6, pp. 4186–4210, 2021, doi: 10.1109/JIOT.2020.3031162.
- [42] B. Tasneem and M. Wahid, "A Review of Secure Routing Challenges in Low Power and Lossy Networks," in *2021 International Conference on Communication Technologies (ComTech)*, Sep. 2021, pp. 120–125, doi: 10.1109/ComTech52583.2021.9616966.
- [43] P. Ruckebusch, J. Devloo, D. Carels, E. De Poorter, and I. Moerman, "An Evaluation of Link Estimation Algorithms for RPL in Dynamic Wireless Sensor Networks," 2016, pp. 349–361.
- [44] P. Sanmartin, A. Rojas, L. Fernandez, K. Avila, D. Jabba, and S. Valle, "Sigma Routing Metric for RPL Protocol," *Sensors*, vol. 18, no. 4, p. 1277, Apr. 2018, doi: 10.3390/s18041277.
- [45] H. Lamaazi and N. Benamar, "OF-EC: A novel energy consumption aware objective function for RPL based on fuzzy logic," *J. Netw. Comput. Appl.*, vol. 117, pp. 42–58, Sep. 2018, doi: 10.1016/j.jnca.2018.05.015.
- [46] X. Liu, Z. Sheng, C. Yin, F. Ali, and D. Roggen, "Performance Analysis of Routing Protocol for Low Power and Lossy Networks (RPL) in Large Scale Networks," *IEEE Internet Things J.*, vol. 4, no. 6, pp. 2172–2185, Dec. 2017, doi: 10.1109/JIOT.2017.2755980.
- [47] O. Gaddour, A. Koubâa, and M. Abid, "Quality-of-service aware routing for static and mobile IPv6-based low-power and lossy sensor networks using RPL," *Ad Hoc Networks*, vol. 33, pp. 233–256, Oct. 2015, doi: 10.1016/j.adhoc.2015.05.009.
- [48] Z. A. Almusaylim, N. Jhanjhi, and A. Alhumam, "Detection and Mitigation of RPL Rank and Version Number Attacks in the Internet of Things: SRPL-RP," *Sensors*, vol. 20, no. 21, p. 5997, Oct. 2020, doi: 10.3390/s20215997.
- [49] U. Shafique, A. Khan, A. Rehman, F. Bashir, and M. Alam, "Detection of rank attack in routing protocol for Low Power and Lossy Networks," *Ann. Telecommun.*, vol. 73, no. 7–8, pp. 429–438, Aug. 2018, doi: 10.1007/s12243-018-0645-4.
- [50] M. A. Boudouaia, A. Abouaissa, A. Ali - Pacha, A. Benayache, and P. Lorenz, "RPL rank based - attack mitigation scheme in IoT environment," *Int. J. Commun. Syst.*, vol. 34, no. 13, Sep. 2021, doi: 10.1002/dac.4917.
- [51] D. S. Nair and S. K. BJ, "Identifying Rank Attacks and Alert Application in WSN," in *2021 6th International Conference on Communication and Electronics Systems (ICCES)*, Jul. 2021, pp. 798–802, doi: 10.1109/ICCES51350.2021.9489034.
- [52] S. Karmakar, J. Sengupta, and S. Das Bit, "LEADER: Low Overhead Rank Attack Detection for Securing RPL based IoT," in *2021 International Conference on COMMunication Systems & NETWORKS (COMSNETS)*, Jan. 2021, pp. 429–437, doi: 10.1109/COMSNETS51098.2021.9352937.
- [53] A. R. Zaroor, "Securing RPL Routing Path for IoT against rank attack via utilizing layering technique," *Int. J. Electr. Eng. Informatics*, vol. 13, no. 4, pp. 789–800, 2021, doi: 10.15676/ijeei.2021.13.4.2.
- [54] R. Stephen and L. Arockiam, "E2V: Techniques for Detecting and Mitigating Rank Inconsistency Attack (RInA) in RPL based Internet of Things," *J. Phys. Conf. Ser.*, vol. 1142, no. 1, 2018, doi: 10.1088/1742-6596/1142/1/012009.
- [55] A. D. Seth, S. Biswas, and A. K. Dhar, "Detection and Verification of Decreased Rank Attack using Round-Trip Times in RPL-Based 6LoWPAN Networks," *Int. Symp. Adv. Networks Telecommun. Syst. ANTS*, vol. 2020-Decem, pp. 3–8, 2020, doi: 10.1109/ANTS50601.2020.9342754.
- [56] A. Althubaity, T. Gong, K. K. Raymond, M. Nixon, R. Ammar, and S. Han, "Specification-based Distributed Detection of Rank-related Attacks in RPL-based Resource-Constrained Real-Time Wireless Networks," *Proc. - 2020 IEEE Conf. Ind. Cyberphysical Syst. ICPS 2020*, pp. 168–175, 2020, doi: 10.1109/ICPS48405.2020.9274726.

Validation of Evacuation Assessment Algorithm in Finding the Best Indoor Evacuation Model

Amir Haikal Abdul Halim, Khyrina Airin Fariza Abu Samah*

Faculty of Computer and Mathematical Sciences
Universiti Teknologi MARA Cawangan Melaka Kampus Jasin, Melaka, Malaysia

Abstract—This paper proposed an indoor evacuation assessment algorithm. Indoor evacuation wayfinding to the nearest exit becomes more difficult due to the intricacy of the inside layout and the involvement of numerous people. Thus, evacuation models were developed by researchers to assist evacuees in safely exiting a building. Unfortunately, building owners are unsure which evacuation model is best for their high-rise buildings. Therefore, we proposed an assessment algorithm to help the owners assess the best evacuation model. This research uses floor plan levels 13 and 14 of Yayasan Melaka's, an office building, to simulate the evacuation. Ten simulation studies for each level are created. The proposed assessment algorithm focuses on three Microscopic evacuation models; agent-based, cellular automata, and social force. Hence, three simulation software were used to represent the mentioned evacuation model: Pathfinder, PedGo, and AnyLogic. K-Mean is then used to cluster the simulation time results. Elbow, Silhouette and V-measure techniques were applied to produce accurate results of the K-Mean. We compiled and analyzed the results from ten simulation studies for each level. The validation was done by comparing the final results. It shows that 70% of the lowest time taken is from Pathfinder, 30% from PedGo, and 0% from AnyLogic. Based on the result, it is proven that the proposed assessment algorithm can provide the best indoor evacuation model followed the attributes set for the building.

Keywords—Assessment algorithm; evacuation model; indoor evacuation; k-mean; validation

I. INTRODUCTION

Evacuation is the organized, regulated, and supervised retreat, dispersal, or withdrawal of individuals from places of risk or hazard and their reception and treatment in secure environments [1]. Despite the limited space available in urban regions, the population of large and medium-sized cities worldwide continues to grow. As a result of the requirement to deal with this development, high-rise buildings have popped up fastly [2]. Thus, fires in high-rise buildings have become more prevalent in recent decades as high-rise structures significantly affect the skylines of major cities [3]. Therefore, proper emergency evacuation in any high-rise structure is critical.

According to the Fire & Rescue Service Department and the Occupational, Health and Safety Environment, the evacuation method by occupants in one building should be able to escape the building 3 minutes after the emergency alarm goes off. Building evacuation must be evaluated for time optimization to avoid human casualties [4]. Evacuees with a misperception of the building environment may display significant rounding or even be trapped, resulting in a

significantly longer evacuation time. According to Ventura [1], people usually take a path of self-estimated speedy escape depending on their current condition. In addition, panic and stamping can lead to several people departing in an emergency. The architecture of escape routes from structures, human psychology and behaviour, and various social and behavioural patterns can significantly influence evacuation performance, resulting in a trapped situation [5]. For instance, a case in Gujarat, India, sacrificed 20 students in a fire because no safety equipment was installed in the building, and there were no escape routes [6]. Another example of disaster is the World Trade Centre (WTC) Twin Towers terrorist attack on September 11, 2001, where 3000 innocent people died [7]. Thus, a high-rise building must have an evacuation strategy to allow evacuees to evacuate the building safely.

Jiang et al. [8] stated there are three types of evacuation models which is microscopic, macroscopic, and mesoscopic. Individuals' geographical and chronological activities are frequently defined by microscopic models [9]. The continuum model, often known as the macroscopic model, integrates variables and monitors characteristics [10]. Finally, mesoscopic models, which focus on groups but offer more specific information about each pedestrian, considered the individuals but not individuals' interactions. The goal is to keep some control over the individual while moving the group as a whole and avoiding local interactions [11]. As a result, mesoscopic is not taken into account in this study. Shi et al. [12] claimed that microscopic and macroscopic models are often used in evacuation evaluations to illustrate pedestrian traffic. Macroscopic models, which reflect overall population movement, do not typically characterize individuals.

On the other hand, microscopic models focus on the smallest of individuals' details. Microscopic models have been employed extensively in recent years [13] in various crowd simulation studies to understand better crowd behaviour in emergency scenarios [14]. For microscopic models, researchers have mostly employed these three models: Agent-based model (ABM), cellular automata (CA), and social force model (SFM) [15]. Thus, the microscopic model is the best among the three types of evacuation models for the indoor evacuation model.

Therefore, this research proposes an intelligent indoor evacuation assessment algorithm for critical incidents. The assessment algorithm can help select the best evacuation model for the chosen building. The best model selection is crucial since it depends on the environment and the building's needs. It also includes the evacuees' ability to evacuate safely and quickly. This paper's organization begins with a brief

*Corresponding Author.

introduction in Section 1. Section 2 explains the related work and is followed by the research methodology in Section 3. Section 4 elaborates on the results and discussion on optimal k number, v-measure score, intracluster distance, and chosen lowest time taken results. Finally, Section 5 concludes the study and briefly mentions future enhancement.

II. RELATED WORK AND TECHNIQUES

This section describes the related works in clustering algorithms and techniques related to the study.

A. Related Work

The related works involved in this research include the K-Mean algorithm and finding the optimal k number. In general, the K-mean approach is dependent on the value of k , which must always be provided before any clustering analysis can be performed. Clustering with various k values will provide diverse outcomes [16]. The algorithm in clustering can be a feature, as an example in Fig. 1. Training examples are shown as dots and cluster centroids as crosses, (a) original dataset, (b) random initial cluster centroids, and (c-f) illustration of running two iterations of K-Means.

The closest cluster centroid is allocated to each training sample in each loop. It is demonstrated by “painting” the training samples with the same colour as the cluster centroid to which they have been allocated. Then, for each cluster, the mean of the points assigned to it is shifted from the centroid to the mean of the points assigned to it.

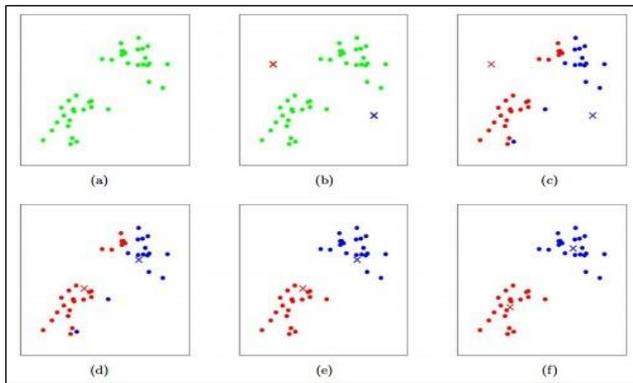


Fig. 1. Clustering example of K-Mean.

The process typically finishes when the centroids stabilize, or the points cease migrating to other groups. However, this depends on the type of grouped data, and the objective function used to quantify proximity. Because K-Mean might have difficulties with local optimum solutions, a proper initialization has proved to be an effective strategy to avoid being caught in the incorrect local optimal solutions [17]. Fig. 2 shows the K-Mean pseudocode [18]. The clustering aims to improve the objective feature (f) by measuring the range between entities and clusters (the most used measurement is the standard Euclidean Distance) as in (1) [19]:

$$f = \sum_{i=1}^K \sum_{j=1}^N \|x_j - C_i\|^2 \quad j \in G_i \quad (1)$$

where K is the number of clusters, N is the number of objects, x_j is the coordinate of object j , C_i is the coordinate of

the cluster i and G_i is the group of objects that belong to cluster i . The algorithm shifts the cluster in space to reduce the square distances within the cluster. The positions of all objects belonging to each cluster are recalculated by averaging. Calculation of the center uses as in (2):

$$C_i = \frac{1}{|G_i|} \sum_{j=1}^N X_j \quad j \in G_i \quad (2)$$

where $|G_i|$ is the number of objects in the cluster i . The algorithm begins with a random set of the C_i cluster’s initial K center points ($i = 1, \dots, K$), which are the present centroids.

<p>Input: K_y: the number of clusters D_y: a data set containing n object</p> <p>Output: A set of K_y clusters</p> <p>Algorithm:</p> <ol style="list-style-type: none"> 1. Input the data set and value of K_y. 2. If $K_y = 1$ then Exit. 3. Else 4. Choose k objects from D randomly as the initial cluster centers. 5. For every data point in the cluster, j reissue and define every object into the cluster where the object matches, based on the object’s mean value. 6. Update cluster means; after that, for each cluster, calculate the object’s mean value. 7. Repeat from step 4 until no data point was assigned; otherwise, stop. <p>The satisfying criteria can be either number of iteration or the change of position of the centroid in consecutive iterations.</p>

Fig. 2. Pseudocode of K-Mean.

Finding the best k number for the cluster is crucial because K-Mean requires a suitable initialization of the k number for clusters to avoid getting trapped at an incorrect local optimal solution. Running the algorithm numerous times and selecting the appropriate number of clusters based on a few validity criteria or automatically identifying them using practical ways or standards is a fundamental way to decide the number of clusters. The process may also change and tweak the cluster centers several times [20]. Several frameworks and techniques have been thoroughly investigated and developed in the past to provide cluster quality measures that indicate if a particular clustering is suitable. There are three ways to verify the clusters, which are called cluster validity index (CVI). These include external, internal, and relative validity indices [21].

More than one index should be used to obtain outstanding and accurate findings [22]. A few methods for determining the best k number have been considered for this study. Two commonly used approaches, the Elbow method and the Silhouette method, are investigated in this study to aid in the manual selection of the number of displayed clusters [23]. Internal validity indexes are used in both methods to assess the correctness of a clustering algorithm [24]. Another technique examined for this study is the V-measure, based on an external validity index. External validity indices such as V-measure are commonly used to determine the best clustering result for a dataset since they know the ‘real’ number of clusters in advance [25], particularly the number of clusters recommended by Elbow and Silhouette techniques for this study. Table I briefly describes the methods used to find the optimal k number for K-Mean.

TABLE I. METHODS TO FIND OPTIMAL *k* NUMBER

Methods	Description	CVI Type
Elbow	The consistency of the optimal number of clusters was visually checked by comparing the difference in each cluster's square error sum (SSE). The best figure is the most significant variation in elbow angle [25].	Internal
Silhouette	Uses a silhouette coefficient that combines separation and coherence. The larger the Silhouette coefficient, the better the cluster [24].	Internal
V-Measure Score	If items in clusters have independent labels, the V-measure is a handy tool for evaluating them. The degree of homogeneity of labels in clusters may be used to measure the quality of clustering objectively [20].	External

B. Related Techniques

The related technique used in this research is the indoor evacuation assessment algorithm based on our previous research [26][27]. Fig. 3 shows the detailed flow of the developed indoor evacuation assessment algorithm. The design and development are separated into six sections in general: 1) determine attributes, 2) run the simulation, 3) identify the best *k* number, 4) evaluate cluster performance, 5) compute intracluster distance, and 6) select the best evacuation model.

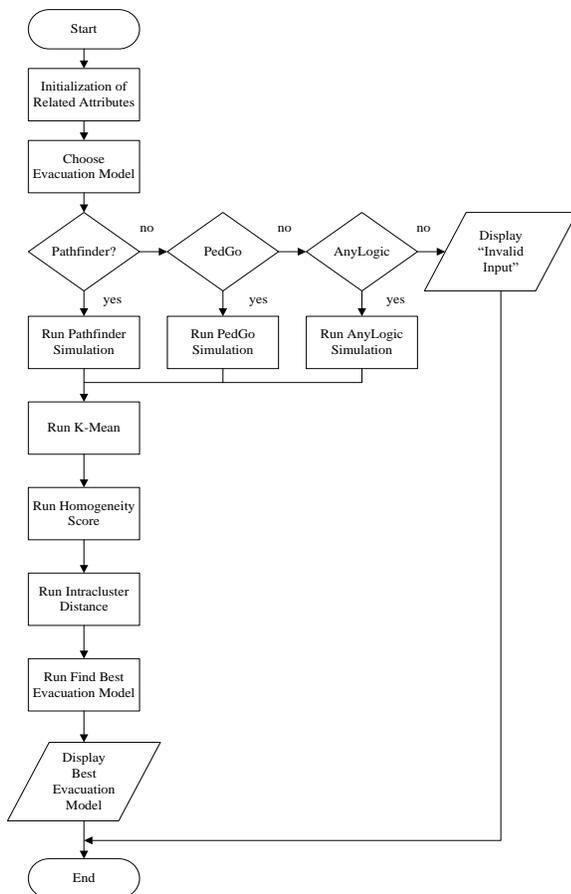


Fig. 3. Overview of Indoor Evacuation Assessment Algorithm.

The attributes involved consist of seven: 1) number of agents, 2) agents' behaviour, 3) room size, 4) number of doors,

5) number of staircases, 6) blockage status, and 7) number of exits. Once entering attribute values, they are used in the selected evacuation simulation software. Three simulation software programs calculate how long the agents will take to evacuate the building. The simulation software involved is Pathfinder, PedGo, and AnyLogic. The K-Mean clustering technique is used once the findings are generated. Several actions are conducted at this point in order to obtain correct findings. The processes involve determining the optimal *k* number, confirming it with the V-measure score, and determining the lowest intracluster distance between clusters to find the lowest time taken. Finally, the assessment algorithm presents the most effective evacuation model.

III. METHODS

This section divides the research methodology into two phases: 1) drawing and mapping floor plans; and 2) simulation studies.

A. Drawing and Mapping Floor Plans

The floor plan of the chosen building is drawn and mapped in the simulation software. The high-rise building used for this research is Yayasan Melaka's building. The chosen floors are levels 13 and 14, which level 14 being the highest level. Yayasan Melaka is a large office with several rooms and barriers that might make evacuation difficult. This construction is a high-rise skyscraper with two access paths on each floor. Staircases are said to be an escape route. Elevators and windows are not permitted to be utilized as exits since elevators are outlawed, and the building's height renders window escape difficult.

The simulation software used to produce time taken results is Pathfinder, PedGo, and AnyLogic. The simulation software represents the evacuation model chosen, ABM, CA, and SFM, respectively. The drawing and mapping of the floor plan are based on the simulation studies created. A few ground rules were observed during the mapping process because each simulation software's functional capabilities vary; such criteria are observed. Two rules are: 1) for each simulation, the paths are set in stone and 2) the agents are positioned in the same room for each simulation software.

As a result, particular simulations require manually mapping the agents' path from the beginning point to the endpoint so that they can travel during the experiment. Fair simulations are ensured by placing agents in the same rooms for each simulation software. The procedures required to map the layouts in each simulation program differ from one another when it comes to mapping.

B. Simulation Studies

The assessment algorithm aims to find the most suitable evacuation model for the given structure. The evacuation simulations were used to apply simulation findings for the research purposes for the assessment process. These simulation studies are implemented in Pathfinder, PedGo, and AnyLogic simulation software. For each level 13 and level 14, ten simulation studies highlight the seven simulation attributes. Level 13 simulation studies are shown in Table II, while level 14 simulation studies are shown in Table III.

TABLE II. SIMULATIONS STUDIES FOR LEVEL 13

Simulation Study	Number of agents	Agents' behaviour	Room size, ft^2	Number of doors	Number of staircases	Blockage condition	Number of exits
SS13-1	50	Group	1.5E	13	26	Yes	2
SS13-2	50	Scattered	1.5E	13	26	Yes	2
SS13-3	100	Group	1.5E	16	26	Yes	1
SS13-4	100	Scattered	1.5E	16	26	Yes	1
SS13-5	150	Group	1.5E	21	26	No	2
SS13-6	150	Scattered	1.5E	21	26	No	2
SS13-7	200	Group	1.5E	20	26	No	1
SS13-8	200	Scattered	1.5E	20	26	No	1
SS13-9	250	Group	1.5E	22	26	Yes	2
SS13-10	250	Scattered	1.5E	22	26	Yes	2

TABLE III. SIMULATIONS STUDIES FOR LEVEL 14

Simulation Study	Number of agents	Agents' behaviour	Room size, ft^2	Number of doors	Number of staircases	Blockage condition	Number of exits
SS14-1	50	Group	1.5E	13	28	Yes	2
SS14-2	50	Scattered	1.5E	13	28	Yes	2
SS14-3	100	Group	1.5E	14	28	Yes	1
SS14-4	100	Scattered	1.5E	14	28	Yes	1
SS14-5	150	Group	1.5E	20	28	No	2
SS14-6	150	Scattered	1.5E	20	28	No	2
SS14-7	200	Group	1.5E	21	28	No	1
SS14-8	200	Scattered	1.5E	21	28	No	1
SS14-9	250	Group	1.5E	22	28	Yes	2
SS14-10	250	Scattered	1.5E	22	28	Yes	2

The values are chosen depending on the building's appropriateness. The number of agents begins at 50 and rises by 50 in each iteration until the total number of agents reaches 250. A group or scattered behaviour distinguishes the agent. The room size is based on the original layout set and is set at $1.5E ft^2$. The number of doors is determined by the total number of doors utilized by the agents, and the number of staircases can either be two or four, depending on the structure. This research uses the time taken for agents to escape using stairs of 0.44m/s for the mean overall movement speed [28], and the length of the stairs is 7384mm from up to down [29]. The requirement for a blockage is assessed, and the number of exits is set to one or two.

IV. RESULT AND DISCUSSION

A. Optimal k number Results

When using K-Mean clustering algorithms, determining the appropriate k number is crucial. The best k number for K-Mean is found using the Elbow and Silhouette approaches. The Elbow and Silhouette method findings and the Silhouette analysis are included in the results. The graph depicts the outcomes of finding the best k number. The elbow point in the graph for the Elbow technique reveals that the point is the ideal k number for determining the optimal k number based on the graphs. The optimum k for the Silhouette technique is the point with the highest silhouette score. The result of visualization

graphs depends on the simulation study; thus, we only show the result for SS13-1 since inserting all the results will take too many pages. Fig. 4 depicts the Elbow method's result where the elbow point can be seen as either 3 or 4. 4 is chosen to be the elbow point. Fig. 5 shows the Silhouette method's result where the highest silhouette score shown is 2. Silhouette analysis in Fig. 6 shows the silhouette plot of the clusters and the visualization of the clustered data. The dotted red line in the silhouette plot of the clusters shows the optimal silhouette coefficient value. Table IV shows the k number results suggested by both Elbow and Silhouette methods for level 13, and Table V shows the k number suggested by both Elbow and Silhouette methods for level 14.

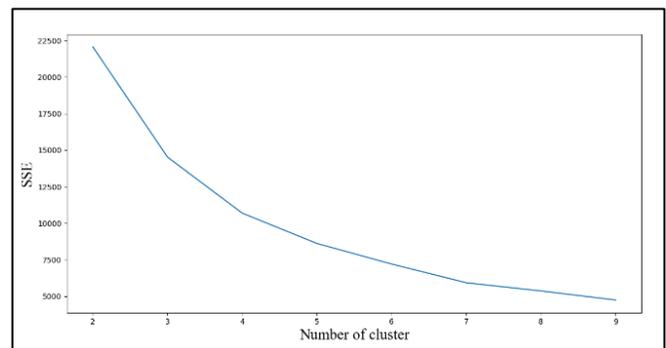


Fig. 4. Elbow method result for SS13-1

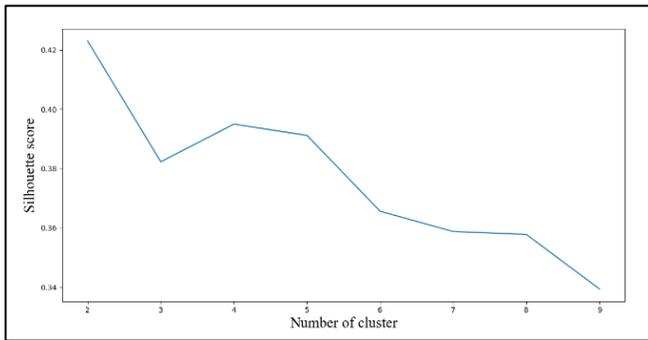


Fig. 5. Silhouette Method Result for SS13-1.

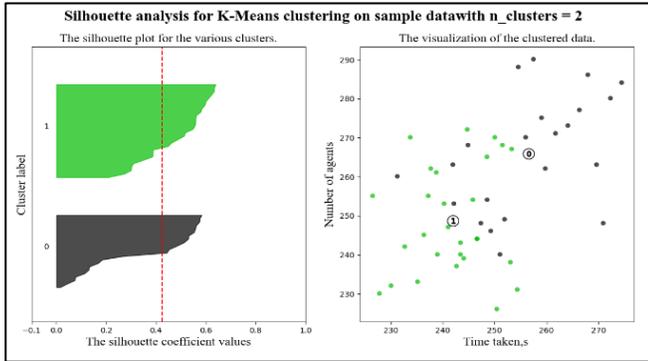


Fig. 6. Silhouette Analysis for SS13-1.

B. V-measure Score Results

The V-measure score is then used to validate the suggested optimal *k* number. It will compare the Elbow and Silhouette methods outcomes. If one of the scores is higher than the other, the Elbow or Silhouette approach with the highest score is picked. Table VI shows the V-measure score results based on the *k* number results suggested by Elbow and Silhouette methods for level 13. Table VII shows the V-measure score

results based on the *k* number results suggested by Elbow and Silhouette methods for level 14. The chosen *k* number is also shown in the tables.

TABLE IV. SUGGESTED OPTIMAL K NUMBERS FOR LEVEL 13

Simulation study	Elbow method	Silhouette method
SS13-1	4	2
SS13-2	5	2
SS13-3	4	2
SS13-4	3	2
SS13-5	3	2
SS13-6	3	2
SS13-7	3	2
SS13-8	3	2
SS13-9	3	3
SS13-10	4	3

TABLE V. SUGGESTED OPTIMAL K NUMBERS FOR LEVEL 14

Simulation study	Elbow method	Silhouette method
SS14-1	-	2
SS14-2	4	4
SS14-3	5	2
SS14-4	3	2
SS14-5	3	2
SS14-6	4	2
SS14-7	4	2
SS14-8	-	2
SS14-9	4	2
SS14-10	4	2

TABLE VI. V-MEASURE SCORE RESULTS FOR LEVEL 13

Simulation study	Elbow method	Silhouette method	Elbow's V-measure Score	Silhouette's V-measure score	Chosen <i>k</i> number
SS13-1	4	2	0.5221779373241466	0.2983631321334766	4
SS13-2	5	2	0.5714202764885019	0.2983631321334766	5
SS13-3	4	2	0.4491895619366153	0.2615824154232080	4
SS13-4	3	2	0.3824680569409242	0.2616480412956257	3
SS13-5	3	2	0.3578833679207950	0.2430208702257761	3
SS13-6	3	2	0.3583568830279575	0.2359561227375162	3
SS13-7	3	2	0.3429001741769688	0.2312453476439503	3
SS13-8	3	2	0.3417016541809229	0.2312453476439503	3
SS13-9	3	3	0.3192609377271065	0.3192609377271065	3
SS13-10	4	3	0.3961601706307684	0.3265114737514012	4

TABLE VII. V-MEASURE SCORE RESULTS FOR LEVEL 14

Simulation study	Elbow method	Silhouette method	Elbow's V-measure Score	Silhouette's V-measure score	Chosen <i>k</i> number
SS14-1	-	2	-	0.3007347242825145	2
SS14-2	4	4	0.4868437889158798	0.4868437889158797	4
SS14-3	5	2	0.5136239262825523	0.2615824154232080	5
SS14-4	3	2	0.3847749621887950	0.2600032659164130	3
SS14-5	3	2	0.3573284764772723	0.2430208702257760	3
SS14-6	4	2	0.4318826415735761	0.2430482521519111	4
SS14-7	4	2	0.4062920631507129	0.2275753301350341	4
SS14-8	-	2	-	0.2311419664100973	2
SS14-9	4	2	0.3964760790239892	0.2215198295727177	4
SS14-10	4	2	0.3953084843355343	0.2228414459888911	4

C. Intracluster Distance Results

The result of time taken from each simulation research is incorporated in K-Mean using the Elbow and Silhouette techniques to discover the optimal *k* number and the V-measure score to decide which optimal *k* number is superior when both approaches are compared. The intracluster distance may then be computed for each cluster in each simulated experiment. The intracluster distance is calculated using Rapidminer. Table VIII shows each simulation study's lowest intracluster distance results for level 13, and Table IX shows each simulation study's lowest intracluster distance results for level 14. The chosen cluster is also shown in the tables.

TABLE VIII. INTRACLUSTER DISTANCE RESULTS FOR LEVEL 13

Simulation Study	Lowest Intracluster Distance	Chosen Cluster
SS13-1	-251.299	3
SS13-2	-181.858	3
SS13-3	-1726.339	3
SS13-4	-1847.029	0
SS13-5	-1265.699	0
SS13-6	-1399.664	0
SS13-7	-3823.050	0
SS13-8	-3723.412	1
SS13-9	-5497.067	2
SS13-10	-4800.702	1

D. Chosen Lowest Time Taken Results

The intracluster distance aids in determining which cluster is ideal for finding the quickest evacuation time. The evacuation model implemented in the chosen building is determined by the lowest time chosen from the three simulation software findings based on each simulation study by level. The simulation software's time-based findings are incorporated into the assessment algorithm, which is then examined and contrasted. For level 13, Table X provides the lowest time taken findings from the selected clusters based on each simulation study and its accompanying simulation software. For level 14, Table XI provides the shortest time taken findings from the selected clusters based on each simulation study and its accompanying simulation software.

TABLE IX. INTRACLUSTER DISTANCE RESULTS FOR LEVEL 14

Simulation Study	Lowest Intracluster Distance	Chosen Cluster
SS14-1	-809.222	0
SS14-2	-323.855	0
SS14-3	-1081.111	3
SS14-4	-1944.537	0
SS14-5	-4780.017	1
SS14-6	-3604.302	2
SS14-7	-1878.660	0
SS14-8	-2626.092	1
SS14-9	-10050.669	2
SS14-10	-10286.633	1

TABLE X. LIST OF LOWEST TIME TAKEN FOR LEVEL 13

Simulation Study	Number of agents	Agents' behaviour	Room size, <i>ft</i> ²	Number of doors	Number of staircases	Blockage Condition	Number of exits	Lowest Time Taken, <i>s</i>	Evacuation Simulation
SS13-1	50	Group	1.5 <i>E</i>	13	26	Yes	2	241.96	Pathfinder
SS13-2	50	Scattered	1.5 <i>E</i>	13	26	Yes	2	243.63	Pathfinder
SS13-3	100	Group	1.5 <i>E</i>	16	26	Yes	1	231.63	Pathfinder
SS13-4	100	Scattered	1.5 <i>E</i>	16	26	Yes	1	231.13	PedGo
SS13-5	150	Group	1.5 <i>E</i>	21	26	No	2	228.26	Pathfinder
SS13-6	150	Scattered	1.5 <i>E</i>	21	26	No	2	227.13	PedGo
SS13-7	200	Group	1.5 <i>E</i>	19	26	No	1	241.26	Pathfinder
SS13-8	200	Scattered	1.5 <i>E</i>	19	26	No	1	235.68	Pathfinder
SS13-9	250	Group	1.5 <i>E</i>	21	26	Yes	2	225.56	Pathfinder
SS13-10	250	Scattered	1.5 <i>E</i>	21	26	Yes	2	250.13	PedGo

TABLE XI. LIST OF LOWEST TIME TAKEN FOR LEVEL 14

Simulation Study	Number of agents	Agents' behaviour	Room size, ft^2	Number of doors	Number of staircases	Blockage Condition	Number of exits	Lowest Time Taken, s	Evacuation Simulation
SS14-1	50	Group	1.5E	13	28	Yes	2	245.92	Pathfinder
SS14-2	50	Scattered	1.5E	13	28	Yes	2	244.12	Pathfinder
SS14-3	100	Group	1.5E	14	28	Yes	1	300.92	PedGo
SS14-4	100	Scattered	1.5E	14	28	Yes	1	256.82	Pathfinder
SS14-5	150	Group	1.5E	20	28	No	2	242.55	Pathfinder
SS14-6	150	Scattered	1.5E	20	28	No	2	242.52	Pathfinder
SS14-7	200	Group	1.5E	21	28	No	1	247.92	PedGo
SS14-8	200	Scattered	1.5E	21	28	No	1	237.95	Pathfinder
SS14-9	250	Group	1.5E	22	28	Yes	2	244.07	Pathfinder
SS14-10	250	Scattered	1.5E	22	28	Yes	2	272.92	PedGo

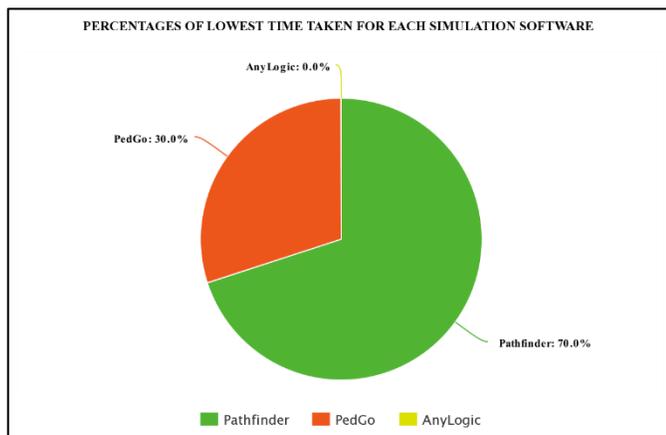


Fig. 7. Piechart of Lowest Time Taken for each Simulation.

As a reminder, the building's architecture determines the appropriateness of existing evacuation models. Different types of evacuation models are suitable for different high-rise structures. Based on the distributed result for each level, Fig. 7 depicts a piechart of the percentages of the lowest time taken for each simulation software. Pathfinder accounts for 70% of the lowest time taken, PedGo for 30%, and AnyLogic for 0%. As a result, it can be determined that ABM is the optimal evacuation model for Yayasan Melaka's building.

V. CONCLUSION

Many developed evacuation models now focus on investigating various evacuation behaviours and times. As a result, the characteristics of the models differ, making it difficult for users to choose a suitable evacuation model. Thus, we presented the indoor evacuation algorithm for high-rise buildings and assessed the proposed solution. Methods and specific attributes for each simulation software were identified, and we managed to compare and analyze the results. It helps to prove how well the assessment algorithm can assess the evacuation model. The result of the lowest time taken has been validated to determine the best evacuation model. Since this study uses a single case study for the simulation and assessment, thus, for future recommendations, the developed assessment algorithm is advised to be evaluated using various high-rise buildings and expand the research by examining the

complete building plan. It is fascinating to compare and contrast because each layout, structure, construction, and fire-resistant capability has its degree of difficulty.

ACKNOWLEDGMENT

The research was funded by Universiti Teknologi MARA Cawangan Melaka under the TEJA Grant 2022 (GDT 2022/1-19).

REFERENCES

- [1] G. M. Ventura, "Patient evacuation resource classification system (PERCS) for residential healthcare facilities: Patient classification system translatable to healthcare evacuation protocols, system modeling, and transportation resources," The George Washington University, 2017.
- [2] N. Ding, T. Chen, Y. Zhu, and Y. Lu, "State-of-the-art high-rise building emergency evacuation behavior," *Physica A: Statistical Mechanics and its Applications*, vol. 561, 2021.
- [3] N. Ding, T. Chen, and H. Zhang, "Simulation of high-rise building evacuation considering fatigue factor based on cellular automata: A case study in China," *Building Simulation*, vol. 10, no. 3, pp. 407–418, 2017.
- [4] H. Sharbini, R. Sallehuddin, and H. Haron, "Crowd evacuation simulation model with soft computing optimization techniques: A systematic literature review," *Journal of Management Analytics*, vol. 8, no. 3, pp. 443–485, 2021.
- [5] Y. Chen, C. Wang, H. Li, J. B. H. Yap, R. Tang, and B. Xu, "Cellular automata model for social forces interaction in building evacuation for sustainable society," *Sustainable Cities and Society*, vol. 53, 2020.
- [6] Z. Siddiqui, "Indian police file case against three over coaching centre fire, death toll rises to 20," 2019.
- [7] L. Zhai, "The comparison of total and phased evacuation strategies for a high-rise office building," 2019.
- [8] Y. Jiang, B. Chen, X. Li, and Z. Ding, "Dynamic navigation field in the social force model for pedestrian evacuation," *Applied Mathematical Modelling*, vol. 80, pp. 815–826, 2020.
- [9] P. Kontou, I. G. Georgoulas, G. A. Trunfio, and G. C. Sirakoulis, "Cellular automata modelling of the movement of people with disabilities during building evacuation," 2018 26th Euromicro International Conference on Parallel, Distributed and Network-based Processing (PDP), pp. 550–557, 2018.
- [10] N. A. A. Bakar, K. Adam, M. A. Majid, and M. Allegra, "A simulation model for crowd evacuation of fire emergency scenario," 2017 8th International Conference on Information Technology (ICIT), pp. 361–368, 2017.
- [11] F. Martinez-Gil, M. Lozano, I. García-Fernández, and F. Fernández, "Modeling, evaluation, and scale on artificial pedestrians: A literature review," *ACM Computing Surveys*, vol. 50, no. 5, 2017.

- [12] M. Shi, E. W. M. Lee, and Y. Ma, "A novel grid-based mesoscopic model for evacuation dynamics," *Physica A: Statistical Mechanics and its Applications*, vol. 497, pp. 198–210, 2018.
- [13] Y. Li, M. Chen, X. Zheng, Z. Dou, and Y. Cheng, "Relationship between behavior aggressiveness and pedestrian dynamics using behavior-based cellular automata model," *Applied Mathematics and Computation*, vol. 371, 2020.
- [14] I. Sakour and H. Hu, "Robot-assisted crowd evacuation under emergency situations: A survey," *Robotics*, vol. 6, no. 2, 2017.
- [15] L. Fayez, "Modeling family behaviours in crowd simulation," Qatar University, 2017.
- [16] M. Ahmed, R. Seraj, and S. M. S. Islam, "The k-means algorithm: A comprehensive survey and performance evaluation," *Electronics (Switzerland)*, vol. 9, no. 8, pp. 1–12, 2020.
- [17] Nurhayati, N. S. Sinatrya, L. K. Wardhani, and Busman, "Analysis of k-means and k-medoids's performance using big data technology," 2018 6th International Conference on Cyber and IT Service Management, CITSM 2018, pp. 1–5, 2019.
- [18] P. Arora, Deepali, and S. Varshney, "Analysis of k-means and k-medoids algorithm for big data," *Physics Procedia*, vol. 78, pp. 507–512, 2016.
- [19] A. B. S. Serapião, G. S. Corrêa, F. B. Gonçalves, and V. O. Carvalho, "Combining k-means and k-harmonic with fish school search algorithm for data clustering task on graphics processing units," *Applied Soft Computing Journal*, vol. 41, pp. 290–304, 2016.
- [20] A. Pugazhenti and L. S. Kumar, "Selection of optimal number of clusters and centroids for k-means and fuzzy c-means clustering: A review," 2020 International Conference on Computing, Communication and Security (ICCCS), pp. 5–8, 2020.
- [21] V. B. B. Anguiano, "Integration and visualization of sparse-grid based clustering methods in the SG++ datamining pipeline," Technical University of Munich, 2019.
- [22] O. Arbelaitz, I. Gurrutxaga, J. Muguerza, J. M. Pérez, and I. Perona, "An extensive comparative study of cluster validity indices," *Pattern Recognition*, vol. 46, no. 1, pp. 243–256, 2013.
- [23] J. Guo, "Developing a visualization tool for unsupervised machine learning techniques on *Omics data," University of Washington, 2018.
- [24] D. M. Saputra, D. Saputra, and L. D. Oswari, "Effect of distance metrics in determining K-Value in KMeans clustering using elbow and silhouette method," *SICONIAN 2019 Sriwijaya International Conference on Information Technology and Its Applications*, vol. 172, pp. 341–346, 2020.
- [25] A. Wani and R. Riyaz, "A new cluster validity index using maximum cluster spread based compactness measure," *International Journal of Intelligent Computing and Cybernetics*, vol. 9, no. 2, pp. 179–204, 2016.
- [26] A. H. A. Halim, K. A. F. A. Samah, Z. Ibrahim, and R. Hamzah, "Conceptual framework for intelligent indoor evacuation model assessment algorithm using integrated assessment model," *International Journal of Advanced Trends in Computer Science and Engineering*, vol. 9, no. 1.4, pp. 289–294, 2020.
- [27] A. H. A. Halim, K. A. F. A. Samah, M. N. H. H. Jono, and L. S. Riza, "A Review on Indoor Evacuation Model and Clustering Techniques in Developing Evacuation Assessment Algorithm," vol. 7, no. 2, pp. 608–619, 2022.
- [28] E. D. Kuligowski et al., "Movement on Stairs During Building Evacuations," National Institute of Standards and Technology Technical Note, no. January, pp. 1–213, 2015.
- [29] H. Hyun-seung, C. Jun-ho, and H. Won-hwa, "Calculating and verifying the staircase-length for evacuation analysis," *Pedestrian and Evacuation Dynamics*, 2011.

Effective Prediction of Software Defects using Random-tree Entropy based Feature Selection Framework

Abdulaziz Alhumam

Department of Computer Science
College of Computer Sciences and Information Technology
King Faisal University, Al-Ahsa, Saudi Arabia

Abstract—Software systems have grown in size and complexity. These characteristics increase the difficulty of preventing software errors. As a result, forecasting the frequency of software module failures is critical to a developer's efficiency. Many methods for defect detection and correcting problems exist. Hence, Machine Learning (ML) classification performance has to be greatly improved. Thus, in this study, a novel approach is proposed for predicting the number of software defects based on relevant variables using ML. First, feature entropy on each raw features is performed and then identifying the un-pruned random feature. Then is selected the relevant feature through the identical existence among the entropy and un-pruned feature. And finally, the software defect dataset of National Aeronautics and Space Administration (NASA) PC-1 is sent to an ML-based model to estimate the number of faults. Initial PC-1 dataset comprises 37 raw features from this only 8 critical characteristics are utilized to enhance the ML model. A random tree feature selection strategy is shown to be accurate and potentially outperform existing methods in the experimental results. The proposed method considerably outperformed the performance of current ML models by obtaining the accuracy of 97.76% in Random Forest (RF) model.

Keywords—Software defect prediction; machine learning; classification; feature entropy

I. INTRODUCTION

In the recent years, the researcher tried to find different techniques and tools in taming the quality, dependability, and reliability of the software systems [1]. A software defect can cause minor inconvenience or catastrophic failure. Pre-deployment fault prediction for testing is supported by recent research in software fault prediction (SFP). Object-oriented programming is harder than procedural programming due to inheritance. By identifying faulty software modules before to the start of the testing process, software defect prediction can help enhance software quality and testing efficiency. These findings aid software engineers in allocating scarce resources to more prone-to-failure modules. Complex software application can deliver high efficient, accurate and powerful work to modern organizations [2]. Software defect prediction (SDP) has grown in popularity during the previous two decades. The results of the SDP assist in allocating resources for software testing. However, defect prediction is often employed for activities with a high degree of precision. It is

difficult to ensure resource allocation prior to software testing or without prior execution data. Machine learning is used to identify problematic modules, as it reveals hidden patterns in software properties [3]. The feature selection activity removes non-classification features with low performance [4]. The variant selection activity selects the best versions of classification methods for their ensemble [5].

A data collection method based on regular expressions and bug-code linking [6] is proposed. In terms of accuracy and consistency, our strategy outperforms other commonly used data collection methods and their publicly available datasets [7]. Around 65 publicly available base datasets containing Chidamber and Kemerer (CK) and other inheritance indicators were used to determine the effect of inheritance on SFP [8]. They investigate the degree to which an inheritance metric accurately predicts software fault proneness. Additionally, they choose CK measures and inheritance metrics for predicting software problems. In SFP experiments, metrics such as exclusive usage and inheritance viability are analyzed [9]. They combed publicly available inheritance metrics data sets and discovered approximately 40 that contained inheritance metrics. Their initial cleanup included nine metrics relating to inheritance.

They preprocessed selected data sets and then merged them using all possible inheritance metrics combinations. The study [10] examined defect prediction datasets. There is no memory data management strategy proposed, nor is a mechanism for defect detection proposed. The proposed technique for defect prediction keeps track of the error rate performance. The defect prediction detector initiates the generation of defects, warning, and control flags. The proposed technique outperforms the conventional technique (p-value 0.05) and within-group comparisons yield statistically significant effect sizes. We observe that increasing the error rate results in DP, which results in suboptimal prediction performance. To overcome the difficulties associated with zero value thresholds, a spectral classifier based on the median absolute deviation threshold was developed [11]. Rather than using a measure of central tendency, this method makes use of the dispersion of eigenvector values. The report's baseline technique is a zero-value threshold spectral classifier, and the entity class is predicted using a heuristic technique.

The highest co-entropy criteria [12] successfully handle the non-Gaussian noise for SDP. A new classifier is created after instance filtering, feature selection, and reduction. It also finds a non-normal distribution for the 21 most significant software indicators. The hybrid feature selection (HFS) [13] is divided into two stages and it clusters features first using hierarchical agglomerative clustering and then eliminates un-normalized and duplicate features using two wrapper methods. Three distinct classifiers with four performance metrics were evaluated empirically on 11 well-studied NASA programs such as accuracy, precision, recall, and F-measure.

II. RELATED WORK

To predict defects on NASA datasets, decision tree (DT), random forest (RF), Naive Bayes (NB), multi-layer perceptron (MLP), radial basis function (RBF), support vector machine (SVM), and k-nearest neighbour classifiers are used [14]. Precision, Recall, F-Measure, Accuracy, Matthew Correlation co-efficient, and ROC Area are used to evaluate classification performance. A two-stage data pre-processing method for software failure prediction models and semi-supervised deep fuzzy C-mean clustering feature extraction is presented [15]. The main goal is to optimise intra-cluster class and feature using deep multi-clusters of unlabelled and labelled data sets. A new strategy called conditional domain adversarial adaptation (CDAA) [16] can help with a variety of SDP problems. The CDAA has a generator, discriminator, and classifier. This is how the generator learns to move between spaces. The discriminator learns to spot the generator's bogus instances. The classifier learns to classify occurrences appropriately. In our CDAA, both classifier and discriminator loss functions propagate to generator. The enhanced wrapper feature selection (EWFS) [17] method selects features in stages while keeping previous choices in mind. This feature selection improves subset assessment while maintaining model performance. On software defect datasets of various granularities, the DT and NB classifiers were used to evaluate EWFS. This feature selection outperformed existing metaheuristics and sequential search-based WFS techniques in the experiments.

For feature exploration and categorization, neural forest (NF) [18] combines deep neural network with decision forest. After the neural network, a decision forest is connected to perform classification and guide feature representation learning. For efficient defect prediction, NF combines NN and decision forests, and the performance of this hybrid method is examined [19]. The hybrid approach [20] improved classification accuracy compared to existing methods. This method investigates the relationship between defect density, velocity, and introduction time. An integrated machine learning approach is used in ten PROMISE data sets with 22838 instances.

To see how FRFS (filter-based ranking feature selection) [21] methods affect software defect using feature selection methods that are too computationally costly. Empirically, they

look at three large-scale web applications. Then they build SVP models using a random forest classifier and seven FRFS methods. To address the prediction model's low classification rates, a hybrid strategy called DELT (diverse ensemble learning technique) [22] is presented. Unlabelled test modules are predicted by majority voting. The DPAHM (Defect Prediction based association hierarchy method) [23] is used to allocate resources for coarse-level activities. FAHP (Fuzzy Analytical Hierarchy Process) is a prevalent multi-criteria decision-making method [24]. Conversely, this evaluation methodology employs a wide range of performance indicators. They may now trust study findings more, avoid misleading conclusions and set realistic restrictions. They employed 11 defect classifiers and 22 prominent performance measurements. The study used KNIME data mining and 12 NASA MDP software defect data sets.

With KMFOS, the class imbalance problem is solved [25]. KMFOS creates additional faulty instances by interpolating between two clusters. They would then spread out in the flawed dataset space. To reduce the noise, CLNI uses cluster-based oversampling. To develop an HDP model, a structured unsupervised deep domain adaptation is applied [26]. They start by combining data from both source and target projects into one statistic. The authors then develop an SNN (simple neural network) model to manage the various and class-imbalanced difficulties in SDP. The hybrid defect prediction model [27] uses the cross-entropy loss function as the classification loss function to reduce distribution mismatch. A heterogeneous defect prediction approach [28], [29] addresses the issue of extreme class imbalance in real-world software datasets. Minority samples in defect data are balanced using the Majority technique based on Mahalanobis distance in the first step. Ensemble learning and joint similarity measurement are used in the second stage to identify the most relevant and representative features across the source and target projects. At last, knowledge transmission from source to target project inside Grassmann manifold space.

The PROMISE Source Code (PSC) dataset was created to expand the CNN research's initial PSC dataset [30]. Our study used 30-repetition holdout and 10-fold cross-validation. An improved CNN model was then proposed and compared to previous CNN findings and an empirical study. It is used to identify contributing elements and independent variables [31]. Defect-free modules have their bugs replaced by a negative number, while faulty modules have their bugs left alone. Negate the false values of defect-free modules while increasing the false values of defective modules. In the next step, algorithms from NASA, SoftLab, and Promise are used. RKEE [32] is preceded by feature selection and rough set-based KNN noise filtering. Remove redundant features first using the feature ranking algorithm. A rough-KNN noise filter removes noisy samples from both minority and majority classes in the second stage. Both the minority and majority classes deal with ambiguity and overlap. NASA and Eclipse data sets have been used to test our technique.

There are considerable discrepancies in data sharing between the source and destination projects, which leads to inconsistencies in metrics. First, we present a clustering-based metric matching approach. An extract multi-granularity metric feature vector unifies the metric dimension while keeping maximum information. A strategy for predicting cross-project defects [33]. That is, it converts the project's original feature space into a manifold space, then uses that manifold space to train a superior naive Bayes prediction model. FSLBDA (few-shot learning based balanced distribution adaptation) technique [34] for unique defect prediction. Under-sampling can correct class imbalance in defect datasets, but reduces the size of training datasets. They remove redundant measurements from severe gradient boosting datasets. Dual innovative approaches [35] for learning from imbalanced data sets to improve minority class forecasting accuracy. These strategies try to distinguish between oversampling and misclassification costs. Experiment findings showed that identifying problematic modules accurately reduced detection system costs by G-mean and AUC. Instance weight is determined by information gravity among source and destination domains, whereas feature load is determined by high correlation with the learning goal, low correlation with other features, and low domain difference. Using 25 real-world datasets, the suggested methodology outperforms existing CPDP (cross project defect prediction) approaches [36]. The suggested approach builds a better CPDP model by allocating weights based on the varying contribution of characteristics and cases to the predictor.

III. PROPOSED METHOD

This section summarizes the software defect classification framework, as well as the significance of each feature. There are a total of 37 software defect attributes in total, with 8 significant features chosen for model performance evaluation. The proposed framework's system block diagram is shown in Fig. 1. NASA software defect datasets must be analyzed using machine learning models. The model is trained using six classification methods in this experiment: DT, EB, RF, SVM, LM, and NN. The experiments are carried out with the help of the R programming language, which trains models to classify software defects. The RF random tree and DT entropy have used feature values for each measurement and measurement class as inputs.

A. Dataset Description

The publicly available NASA Defect Dataset of PC1 was used in this study which is presented in Table I. In the dataset, there are 759 samples and 37 features, respectively. Lines of code, normalised cyclomatic complexity, cyclomatic density, essential complexity, maintenance severity, halstead content, halstead difficulty, parameter count, and other metrics are included in the data as presented in the Table II.

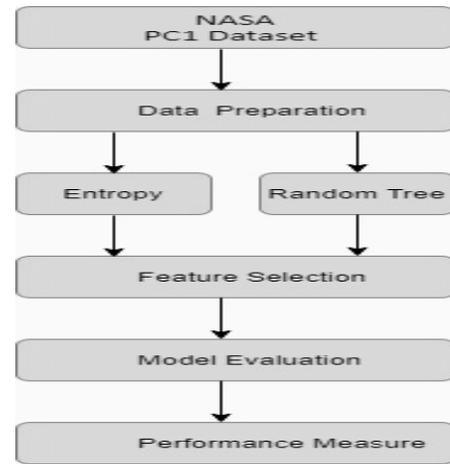


Fig. 1. Experiment Workflow with NASA- PC1.

TABLE I. PC1 FEATURES OF NASA DATA SET

F1_Loc_blank	F19_Halstead_difficulty
F2_Branch_count	F20_Halstead_effort
F3_Call_pairs	F21_Halstead_error_est
F4_Loc_code_and_comment	F22_Halstead_length
F5_Loc_comments	F23_Halstead_level
F6_Condition_count	F24_Halstead_prog_time
F7_Cyclomatic_complexity	F25_Halstead_volume
F8_Cyclomatic_density	F26_Maintenance_severity
F9_Decision_count	F27_Modified_condition_count
F10_Decision_density	F28_Multiple_condition_count
F11_Design_complexity	F29_Node_count
F12_Design_density	F30_Normalized_cyclomatic_complexity
F13_Edge_count	F31_Num_operands
F14_Essential_complexity	F32_Num_operators
F15_Essential_density	F33_Num_unique_operands
F16_Loc_executable	F34_Num_unique_operators
F17_Parameter_count	F35_Number_of_lines
F18_Halstead_content	F36_Percent_comments
	F37_Loc_Total

TABLE II. PC1 FEATURES SELECTED FROM NASA DATA SET

Cyclomatic_density	Halstead_difficulty
Essential_complexity	Maintenance_severity
Parameter_count	Normalized_cyclomatic_complexity
Halstead_content	Number_of_lines

B. Algorithm

The algorithm was input with different dataset of different raw features along with m sample, the different models were trained with, and the performance model was observed for True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) for the model performance evaluation. The steps were followed for each feature with various entropy features using recursive partitioning and with different decision criteria.

-
- Input dataset with n raw features $F_r = \{f_1, f_2, f_3, \dots, f_n\}$ and m samples.
1. Train various models $TM_s = \{tm_1, tm_2, tm_3, tm_s\}$ to observe the TP, TN, FP, FN for model performance evaluation.
 2. **for** each feature **do**
 if \exists a entropy \in raw feature F_r **then**
 Execute entropy features $F_c = \{f_1, f_2, f_3, \dots, f_c\}$ using recursive partitioning and decision criteria.
end if
end for
 3. **for** each feature **do**
 if \exists a significance \in raw feature F_r **then**
 Execute RF variable significance features $RF_v = \{rf_1, rf_2, rf_3, \dots, rf_v\}$ using un-pruned random tree with less error.
end if
end for
 4. Obtain the significant features $SF_k = \{sf_1, sf_2, sf_3, \dots, sf_k\}$ using the Step 2 and 3 as follows:
 Significant features $SF_k = F_c \cap SF_k$
 5. Evaluate the model TM_s performance using significant features SF_k of Step 4.
 6. **for** each model TM_s **do**
 if \exists a model with high accuracy **then**
 Select the model for classification
end if
end for
-

IV. RESULT AND DISCUSSION

In this section, the summary of the experimental results obtained by various machine-learning models are presented. These experiments are conducted on the dataset NASA PC1 Dataset. The results obtained from various ML models are shown in Table III. In the next stage, the ML method on the dataset with all the features of confusion matrix calculated and shown in Fig. 2. The accuracy and precision are also calculated and are shown in Fig. 2 and Fig. 3. The Sensitivity and Specificity are also calculated the results are shown in Fig. 4.

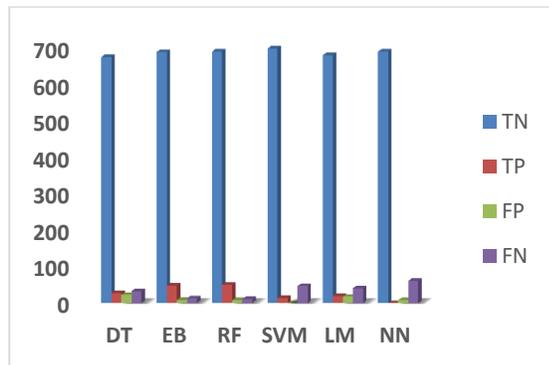


Fig. 2. Confusion Matrix of ML Models.

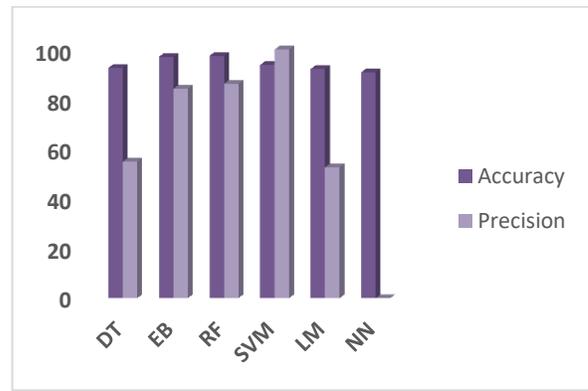


Fig. 3. Accuracy and Precision of ML Models.

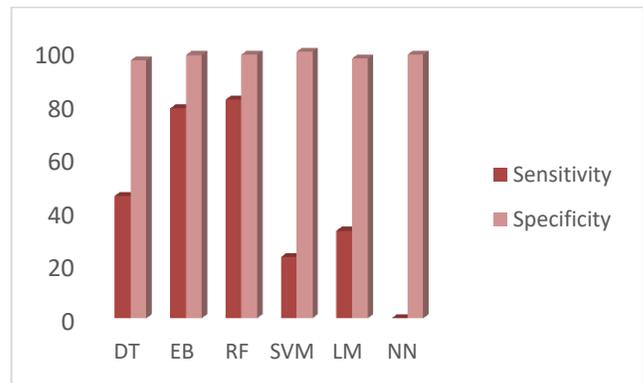


Fig. 4. Sensitivity and Specificity of ML Models.

The results obtained from ML models are shown in Fig. 5.

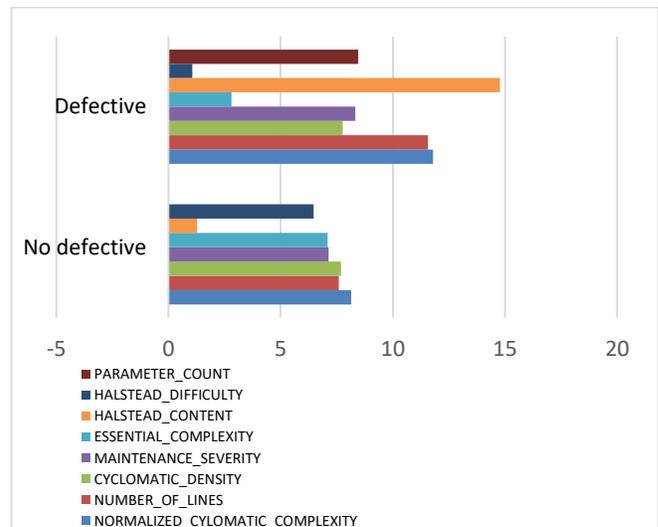


Fig. 5. Significant Features of Defective and Non-defective Class.

In the next stage, the ML method on the dataset with significant features are applied and confusion matrix calculated and shown in Fig. 6. The accuracy and precision with significant features are calculated and the presented in Fig. 7. The Sensitivity and Specificity are also calculated the results are shown in Fig. 8.

TABLE III. RESULTS OF ML MODEL WITHOUT SIGNIFICANT FEATURES

ML	n	TN	TP	FP	FN	Accuracy	Error Rate	Sensitivity	Specificity	Precision
DT	759	675	28	23	33	92.62	7.38	45.90	96.70	54.90
EB	759	689	48	9	13	97.10	2.90	78.69	98.71	84.21
RF	759	690	50	8	11	97.50	2.50	81.97	98.85	86.21
SVM	759	698	14	0	47	93.81	6.19	22.95	100.00	100.00
LM	759	680	20	18	41	92.23	7.77	32.79	97.42	52.63
NN	759	690	0	8	61	90.91	9.09	-	98.85	-

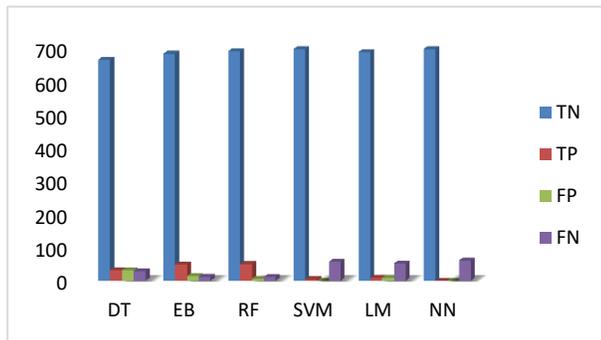


Fig. 6. Confusion Matrix of ML Models with Significant Features.

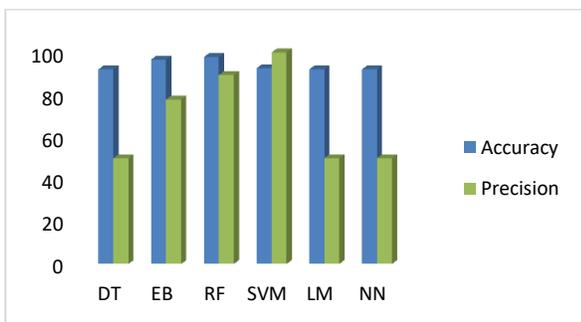


Fig. 7. Accuracy and Precision of ML Models with Significant Features.

Although this result is obtained by using with only 8 features out of 37 features, the proposed approach time consuming due to the large number of parameters in features

selection. Since the proposed model utilizes only important features and avoid features which are not have high impact. Machine Learning models are used in to find out optimal feature selection and significant result improvement achieved by using Random Forest method in selection process.

The results revealed that our proposed method performed better than existing methods without significant features. The machine learning models with all features accuracy results obtained 97.50 % by using Random Forest method. The same dataset with significant features results in accuracy improvement 97.76 is achieved. Six distinct models are investigated for software defect data classification with selected features. As a result, the results of all six classification methods are compared using the outputs of the suggested feature ranking algorithms as input. The experimental results in Table IV shows that the suggested feature with an RF model have the greatest accuracy scores of all six features.

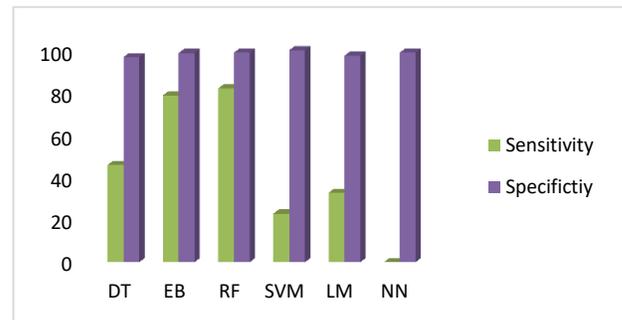


Fig. 8. Sensitivity and Specificity of ML Models with Significant Features.

TABLE IV. RESULTS OF ML MODEL WITH SIGNIFICANT FEATURES

ML	TN	TP	FP	FN	Accuracy	Precision	Sensitivity	Specificity
DT	666	32	32	29	91.96	50.00	52.46	95.42
EB	684	49	14	12	96.57	77.78	80.33	97.99
RF	692	50	6	11	97.76	89.29	81.97	99.14
SVM	698	4	0	57	92.49	100.00	6.56	100.00
LM	689	9	9	52	91.96	50.00	14.75	98.71
NN	698	0.1	0.1	61	91.95	50.00	0.16	99.99

V. CONCLUSION

Software defect prediction method plays important role and important to prevent and predict the bugs in the software in early stages are very difficult and challenging. However, this work using machine learning models perform evaluation of defect prediction with all features used in NASA dataset. The Machine learning models like DT, EB, RF, SVM, LM, and NN are used. The evaluation process carried out using with significant features and all features. The experimental results analyzed and summarized based on confusion matrix, accuracy, precision, sensitivity and specificity. The accuracy is plays major role and error rate also evaluated by using random forest the results are improved. The comparison results with all features and significant features used with ML models shows improvements. As future work, many more ML models and performs comparison among them to make more optimal results.

ACKNOWLEDGMENT

The author wishes to thank the College of Computer Sciences and Information Technology, King Faisal University, Saudi Arabia, for providing the infrastructure for this study.

REFERENCES

- [1] S. K. Alferidah and S. Ahmed, "Automated Software Testing Tools," Proceedings-2020 IEEE, International Conference on Computing and Information Technology, ICCIT-1441, 2020, pp. 1-4.
- [2] A. A. Alsayyah and S. Ahmed, "Energy Efficient Software Development Techniques for Cloud based Applications," International Journal of Advanced Trends in Computer Science and Engineering, vol. 9, no. 5, pp. 8043-8054, 2020.
- [3] D. Chen, X. Chen, H. Li, J. Xie and Y. Mu, "DeepCPDP: Deep learning based cross-project defect prediction," IEEE Access, vol. 7, pp. 184832-184848, 2019.
- [4] U. Ali, S. Aftab, A. Iqbal, Z. Nawaz, M. S. Bashir et al., "Software Defect Prediction Using Variant based Ensemble Learning and Feature Selection Techniques," International Journal of Modern Education and Computer Science, vol. 12, no. 5, pp. 29-40, 2020.
- [5] H. Alsawalqah, N. Hijazi, M. Eshtay, H. Faris, A. A. Radaideh et al., "Software Defect Prediction Using Heterogeneous Ensemble Classification Based on Segmented Patterns," Applied Science, vol. 10, no. 1745, 2020.
- [6] G. Mauša, T.G. Grbac and B.D. Bašić, "A systematic data collection procedure for software defect prediction," Computer Science and Information Systems, vol. 13, no. 1, pp. 173-197, 2016.
- [7] M. A. Alshammari and M. Alshayeb, "The effect of the dataset size on the accuracy of software defect prediction models: An empirical study," Inteligencia Artificial, vol. 24, no. 68, pp. 72-88, 2021.
- [8] S. R. Aziz, T. Khan and A. Nadeem, "Experimental validation of inheritance metrics' impact on software fault prediction," IEEE Access, vol. 7, no. 8742643, pp. 85262-85275, 2019.
- [9] S.R. Aziz, T. A. Khan and A. Nadeem, "Exclusive use and Evaluation of Inheritance Metrics Viability in Software Fault Prediction—An Experimental Study," Peer. J Computer Science, 7, pp. 1-47, 2021.
- [10] M. A. Kabir, J. W. Keung, K. E. Bennin and M. Zhang, "A Drift Propensity Detection Technique to Improve the Performance for Cross-Version Software Defect Prediction," Proceedings - 2020 IEEE 44th Annual Computers, Software, and Applications Conference, COMPSAC 2020, art. no. 9202527, pp. 882-891.
- [11] A. Marjuni, T. B. Adji and R. Ferdiana, "Unsupervised software defect prediction using median absolute deviation threshold based spectral classifier on signed Laplacian matrix," Journal of Big Data, vol. 6, no. 1, 2019.
- [12] H. Ji and S. Huang, "A New Framework Consisted of Data Preprocessing and Classifier Modelling for Software Defect Prediction," Mathematical Problems in Engineering, no. 9616938, 2018.
- [13] Y. Jian, X. Yu, Z. Xu and Z. Ma, "A hybrid feature selection method for software fault prediction," IEICE Transactions on Information and Systems, vol. 10, pp. 1966-1975, 2019.
- [14] A. Iqbal, S. Aftab, U. Ali, Z. Nawaz, L. Sana et al., "Performance analysis of machine learning techniques on software defect prediction using NASA datasets," International Journal of Advanced Computer Science and Applications, vol. 10, no. 5, pp. 300-308, 2019.
- [15] A. Arshad, S. Riaz, L. Jiao and A. Murthy, "The empirical study of semi-supervised deep fuzzy c-mean clustering for software fault prediction," IEEE Access, vol. 6, no. 8439927, pp. 47047-47061, 2018.
- [16] L. Gong, S. Jiang and L. Jiang, "Conditional Domain Adversarial Adaptation for Heterogeneous Defect Prediction," IEEE Access, vol. 8, no. 9169630, pp. 150738-150749, 2020.
- [17] A.O. Balogun, S. Basri, L.F. Capretz, S. Mahamad, A. A. Imam et al., "Software defect prediction using wrapper feature selection based on dynamic re-ranking strategy," Symmetry, vol. 13, no. 11, 2021.
- [18] Y. Qiu, Y. Liu, A. Liu, J. Zhu and J. Xu, "Automatic Feature Exploration and an Application in Defect Prediction," IEEE Access, vol. 7, no. 8794540, pp. 112097-112112, 2019.
- [19] E. A. Felix and S.P. Lee, "Integrated Approach to Software Defect Prediction," IEEE Access, vol. 5, no. 8058420, pp. 21524-21547, 2017.
- [20] M. Banga, A. Bansal and A. Singh, "Proposed hybrid approach to predict software fault detection," International Journal of Performability Engineering, vol. 15, no. 8, pp. 2049-2061, 2019.
- [21] X. Chen, Z. Yuan, Z. Cui, D. Zhang and X. Ju, "Empirical studies on the impact of filter-based ranking feature selection on security vulnerability prediction," IET Software, vol. 15, no. 1, pp. 75-89, 2021.
- [22] U. S. Bhutamapuram and R. Sadam, "With-in-project defect prediction using bootstrap aggregation based diverse ensemble learning technique," Journal of King Saud University - Computer and Information Sciences, (In Press), 2021.
- [23] C. Cui, B. Liu, P. Xiao and S. Wang, "Can Defect Prediction Be Useful for Coarse-Level Tasks of Software Testing?," Applied Sciences, vol. 10, no. 15, 2020.
- [24] H. Ghunaim and J. Dichter, "Applying the FAHP to Improve the Performance Evaluation Reliability of Software Defect Classifiers," IEEE Access, vol. 7, no. 8710236, pp. 62794-62804, 2019.
- [25] L. Gong, S. Jiang and L. Jiang, "Tackling Class Imbalance Problem in Software Defect Prediction through Cluster-Based Over-Sampling with Filtering," IEEE Access, vol. 7, no. 8861051, pp. 145725-145737, 2019.
- [26] L. Gong, S. Jiang, Q. Yu and L. Jiang, "Unsupervised deep domain adaptation for heterogeneous defect prediction," IEICE Transactions on Information and Systems, E102D, pp. 537-549, 2019.
- [27] Y. Shao, J. Zhao, X. Wang, W. Wu and J. Fang, "Research on Cross-Company Defect Prediction Method to Improve Software Security," Security and Communication Networks, no. 5558561, 2021.
- [28] K. Jiang, Y. Zhang, H. Wu, A. Wang and Y. Iwahori, "Heterogeneous defect prediction based on transfer learning to handle extreme imbalance," Applied Sciences, vol. 10, no. 1, 2020.
- [29] A. Wang, Y. Zhang and Y. Yan, "Heterogeneous Defect Prediction Based on Federated Transfer Learning via Knowledge Distillation," IEEE Access, vol. 9, no. 9352701, pp. 29530-29540, 2021.
- [30] C. Pan, M. Lu, B. Xu and H. Gao, "An improved CNN model for within-project software defect prediction," Applied Sciences, vol. 9, no. 10, 2019.
- [31] J.-H. Ren and F. Liu, "Predicting software defects using self-organizing data mining," IEEE Access, vol. 7, no. 8758097, pp. 122796-122810, 2019.
- [32] S. Riaz, A. Arshad and L. Jiao, "Rough Noise-Filtered Easy Ensemble for Software Fault Prediction," IEEE Access, vol. 6, no. 8435900, pp. 46886-46899, 2018.

- [33] Y. Sun, X-Y. Jing, F. Wu, and Y.Sun, "Manifold embedded distribution adaptation for cross-project defect prediction," *IET Software*, vol. 14, no. 7, pp. 825-838, 2020.
- [34] A. Wang, Y. Zhang, H. Wu, K. Jiang and M. Wang, "Few-Shot Learning Based Balanced Distribution Adaptation for Heterogeneous Defect Prediction," *IEEE Access*, vol. 8, no. 8999527, pp. 32989-33001, 2020.
- [35] J. Zheng, X. Wang, D. Wei, B. Chen and Y. Shao, "A Novel Imbalanced Ensemble Learning in Software Defect Predication," *IEEE Access*, vol. 9, no. 9404009, pp. 86855-86868, 2021.
- [36] Q. Zou, L. Lu, S. Qiu, X. Gu and Z. Cai, "Correlation feature and instance weights transfer learning for cross project software defect prediction," *IET Software*, vol.15, no. 1, pp. 55-74, 2021.

Parameter Optimization of Nonlinear Piezoelectric Energy Harvesting System for IoT Applications

Li Wah Thong¹

Faculty of Engineering and Technology
Multimedia University, Melaka, Malaysia

Swee Leong Kok², Roszaidi Ramlan³

Faculty of Electronics and Computer Engineering
Universiti Teknikal Malaysia Melaka, Melaka, Malaysia

Abstract—The vibrational energy harvesting has been essentially applied to power up low-power electronics, microsystems, and wireless sensors especially in the areas of Internet of Things (IoT) devices. This paper investigates the prospect of incorporating nonlinearity in a unimorph piezoelectric cantilever beam with a tip magnet placed under a harmonic base excitation in IoT enabled environment. An empirical and theoretical analysis on the impact of various parameters such as spacing distance between magnets, presence of magnetic tip mass and positioning of vibrational source on the frequency response output was performed. It was observed that the largest spectrum of frequency can be produced when at the lowest resonant frequency of the cantilever. The positioning of vibrational source deeply impacts the hysteresis region and frequency range in realizing broadband energy harvesting. The inclusiveness of vibration source on both the cantilever beam as well as the external magnets impacts the energy harvester in terms of frequency range and the minimal distance for bistable condition.

Keywords—Energy harvesting; nonlinear dynamics; piezoelectric; vibration; broadband frequency

I. INTRODUCTION

The innovations in low power electronics for Internet of Things (IoT) in a variety of capacities ranging from health facilities, smart homes, and security surveillance has been extensively researched in the literatures [1, 2]. These low power sensors and devices traditionally function using batteries that have a limited lifespan. As the demand of sustainability in IoT devices increases, the need to recharge or replace these conventional batteries are essential to ensure the system is fully operational thoroughly [3]. The application of piezoelectric cantilever effect in transforming surrounding mechanical vibration energy to operational electrical energy has gained significant research interest progressively over the years. The vibrational energy harvesting has been applied widely to supply electrical energy to power up low-power electronics, microsystems, and wireless sensors [4-6]. It has also been regarded as an alternative source to chemical batteries that are relatively small and have restricted life duration. Piezoelectric energy harvesters are usually integrated in areas where the usage of batteries is unfeasible and inappropriate. The most commonly deliberated vibration-to-electrical energy conversion mechanism in the literature is modeled using either electromagnetic, electrostatic, magnetostrictive, triboelectric or piezoelectric energy mechanism [7-9].

The single cantilever piezoelectric energy harvester (PEH) that was projected irrespective of load as a linear resonator does not provide efficient operating frequency as it characteristically suffers from a constricted operating frequency range [10]. The simplicity in incorporating piezoelectric harvesters has encouraged in-depth research to increase its performance and extending its operational frequency spectrum. Since piezoelectric energy harvester is a resonant-based model, one of the requirements to obtain maximum electrical power output is to ensure that the excitation frequency of the ambient surroundings matches to the resonant frequency of the harvester [11, 12]. A minor frequency disparity or deviations from the harvester resonant frequency will significantly reduce the harvested power as the imparted stress-strain effect in the piezoelectric mechanism has been affected accordingly. Unfortunately, in real world circumstances, the frequency spectrum of the vibrational energy in ambient environment changes dynamically and can be unpredictable over time [13]. In order to mitigate this issue, evolutions in design and parameters of the piezoelectric energy harvesting system has been expanded in several aspects such as altering design configurations [14, 15], manipulating mechanical nonlinearities [16, 17], and improving electronic circuitry [18, 19] with intentions to increase the spectrum of operational frequency and thus its power output. These innovations have seen viable success in its intended purpose to provide better solutions for vibrational energy harvesters [20].

In view of the design aspects discussed in the literatures, most of the methodology focuses on either tuning the resonant frequency or broadening the operational frequency spectrum of the energy harvester [21, 22]. For techniques involving active resonant frequency tuning, the resultant development of the energy harvester may not be practicable as the tuning of the actuators actually consume more power than the device can harvest [23, 24]. In passive resonant tuning, the increase usage of sensors and actuators indirectly create an upsurge in cost and intricacy of the system [25]. Current research for increasing the bandwidth of the operating frequency has developed positive prospective in recent years in terms of extending range of frequencies to be processed [26-29]. This innovative strategy allows the energy harvester to response to several vibration frequency excitations at the same time. There is no tuning mechanism involved; however, there may be a decrease of maximum power harvested at the instance. The strategies involved in widening the operational frequency spectrum includes designing an array of piezoelectric cantilevers as generators [30], introducing mechanical stopper to limit the

This work is sponsored by the Ministry of Higher Education Malaysia under Fundamental Research Grant scheme (FRGS/1/2020/TK0/MMU/03/13)

harvester amplitude [31], and accustoming nonlinearities into the harvester system through bistable configurations [32, 33]. Many researchers have enthusiastically expanded the bistable configurations which comprises of two stable equilibria energy states created by the exploitations of magnetic field force [34, 35]. Further exploration in the effect of restoring force by magnets and its nonlinearity may create the opportunity to extend the bandwidth of operating frequency in the energy harvesting system. With these motivations, the empirical study in this research paper thus focuses on widening the operational frequency spectrum through exploitations of magnetic field. In the literatures, most researchers study the impact of permanent magnets when only the piezoelectric cantilever was placed under seismic vibration while the external magnets were assumed to be static [36-39]. Some researchers also studied the impact of permanent magnets when both the piezoelectric cantilever beam and external magnet are placed under the same vibration. As the beam and its external magnet of the piezoelectric energy harvester are usually designed inclusively and placed together on the same vibrating platform, both the beam and external magnets will experience the same vibration and will counteract during vibration. Thus, the need to analyse and model a nonlinear energy harvesting system based on the positioning of vibration source is essential for further improvement of broadband energy harvesting.

In this paper, the noteworthy impact in terms of resonant frequency and the voltage output of the piezoelectric energy harvesting system as the relative distance between two fixed magnets in repulsion mode varies horizontally was investigated. The proposed model is designed by incorporating magnets as mass on the single unimorph piezoelectric cantilever beam and inducing the effects of magnetic field force to control the stress-strain effect in the cantilever beam. This paper discusses an experimental analysis on the effect of using magnets and its variations in spacing range towards the bandwidth of frequency spectrum for a wideband bistable energy harvesting system. Furthermore, the influence of positioning the vibration source on either the cantilever beam or both the beam and external magnets are compared to analyse its effect on the resonance frequency and voltage output of the energy harvester. A comparison in term of resonance frequency, output voltage and its spectrum of frequency between both scenarios will be done accordingly.

This paper is organized as follows. Firstly, the research methodology and its concept in using the magnets as load for the piezoelectric cantilever are described in Section 2. Furthermore, Section 3 presents the consequence of extending the horizontal displacement between the two magnets on the piezoelectric resonant frequency. In Section 4, further investigation is done to observe the consequential outcome of spacing distance between magnets on the performance of the harvester. Lastly, a summary of the overall research results in the paper is established and elaborated in the final section.

II. THEORETICAL ANALYSIS

A. Modeling and Design Analysis

The design of the bistable piezoelectric energy harvesting system involves a unimorph piezoelectric cantilever beam and two magnets that will be applied as mass and also for its

repulsive-attractive mechanism. The piezoelectric cantilever is setup by clamping one of its end firmly onto the vibration shaker with the aim to decrease the effect of gravity. At the other free end of the cantilever beam, a magnetic mass that weighs approximately 0.75 gram will be fixed to the piezoelectric cantilever to behave as a mass as well as to be responsible for providing the magnetic force restoration for the system. Subsequently, an additional alike magnetic mass is secured in a fixed location but is positioned in reverse polarity compared with the magnetic mass on the cantilever free end. During the experiment, the magnet in the fixed position will be displaced along the x -axis accordingly to provide displacement, d . The setup of these magnets will provide a variation of repulsive magnetic force strength as the distance between both magnets adjusted accordingly. The controllable vibration shaker functions to provide a transverse harmonic displacement for the piezoelectric cantilever throughout the experiment. In this paper, a comparison will be done to observe the impact of the magnets on the resonance frequency and voltage output when the vibration source is placed differently as illustrated in Fig. 1. In Fig. 1(a), the piezoelectric cantilever beam is placed on the vibration shaker and the magnetic mass is fixed to a stationary structure, namely piezoelectric beam with stationary magnet (PSM) system. While in Fig. 1(b), both the piezoelectric cantilever beam and the fixed magnet will be placed under the same vibration source applying the same base excitations, namely piezoelectric cantilever under the same vibration (PVM) system.

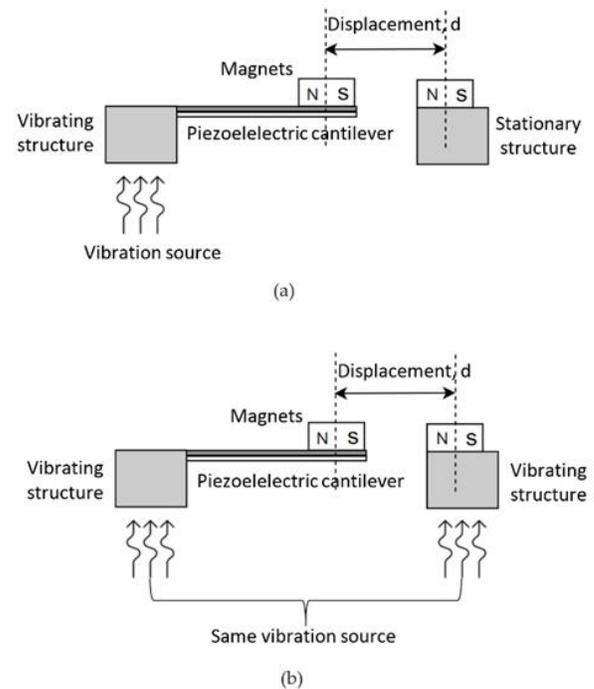


Fig. 1. Setup of Piezoelectric Energy Harvester for (a) PSM (b) PVM.

In electromechanical modeling, a linear resonant piezoelectric harvester can be represented based on Euler-Bernoulli beam equation. In Fig. 1, if the magnets are taken out and replaced with a similar sized mass, the piezoelectric harvester becomes a linear system that can be described using the governing electromechanical model, as in (1) and (2).

$$M\dot{x}'(t) + Cx'(t) + Kx(t) - \theta v(t) = F(t) \quad (1)$$

$$C_p \dot{v}(t) + \frac{v(t)}{R} + \theta \dot{x}(t) = 0 \quad (2)$$

Where M and C denotes the total mass and damping experienced by the energy harvester respectively. K and θ characterize the effective stiffness and the equivalent linear piezoelectric electromechanical coupling coefficient. C_p characterizes the corresponding capacitance of the piezoelectric substance and $v(t)$ denotes the voltage across the external load resistance, R . The vertical tip displacement of the mass is represented by $x(t)$ while $F(t)$ is force mechanically induced by the surrounding vibration excitation [40-42].

When the proof mass of the energy harvester is interchanged with magnets and arranged in repulsive mode, a magnetic repulsive force, F_m exists between the magnets and eventually changes the linear system defined by (1) and (2) to be magnetically coupled non-linear piezoelectric harvester. The non-linear energy harvesting system can then be described, as in (3) and (4).

$$M\ddot{x}(t) + C\dot{x}(t) + Kx(t) - \theta v(t) = F(t) + F_m \quad (3)$$

$$C_p \dot{v}(t) + \frac{v(t)}{R} + \theta \dot{x}(t) = 0 \quad (4)$$

Where the nonlinear magnetic force, F_m can be articulated as the polynomial equation of.

$$F_m = \mu x(t) + \lambda x^3(t) \quad (5)$$

B. Effect of Distance on Interaction between Magnets

Since the stiffness of the piezoelectric cantilever beam is dependent on the magnetic force in the system, the position and distance between magnets will also be one of the factors that will determine the energy harvester performance. In order to create bistability, two magnets are placed in opposite polarity with a specific distance, d along the x -axis of the beam as illustrated in Fig. 2. Due to the magnetic repulsive force, F_m amid the magnets, the actual bending of the beam tip can be characterized as the vertical displacement, h of the tip mass.

Theoretically, a magnetic repulsive force, F_m exists between the magnets and its magnitude will decrease as the distance, d between the magnet increases. In order to observe the impact of this repulsive force on the system, a simplified lumped parameter model is used to represent the physical system as well as to approximate the effect of distance to the respective force as shown in Fig. 3.

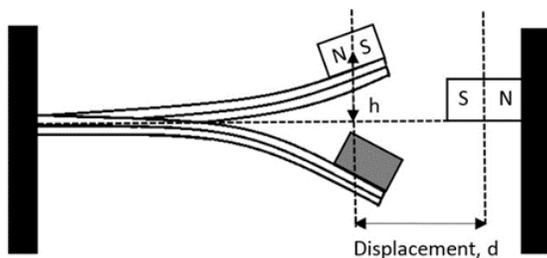


Fig. 2. A Piezoelectric System with Adjustable Distance between Magnets.

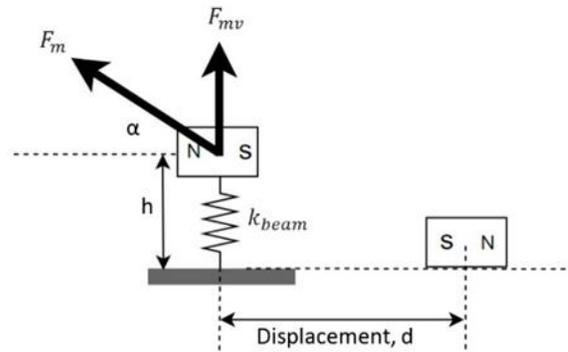


Fig. 3. A Simplified Lumped Parameter Model for the Piezoelectric System.

The total mass, m of the energy harvester system is determined by the effective mass of the first flexural mode of the beam plus the magnetic mass, while the effective spring stiffness, k denotes the elastic response of the piezoelectric cantilever beam [40]. The repulsive force, F_m is assumed to have a constant magnitude for a given value of distance, d provided that the angle θ is reasonably small. Thus, as the mass on the cantilever tip moves, the repulsive force, F_m changes in the direction by an angle α . Since the longitudinal stiffness of the piezoelectric cantilever is assumed to be adequately high, the horizontal component of the repulsive force, F_m will be balanced off. The vertical component of the repulsive force, F_{mv} which affects the stiffness and motion of the energy harvester can be described as.

$$F_{mv} = F_m \sin \alpha \quad (6)$$

The distance, d between the magnets and its relation to the vertical mass displacement, h can be written as.

$$h = d \tan \alpha \quad (7)$$

Consequently, the resultant relationship between the distance, d and the vertical constituent of the repulsive force, F_{mv} can be further deduced as.

$$F_{mv} = F_m \sin \alpha = F_m \frac{h}{\sqrt{d^2 + h^2}} \quad (8)$$

Applying Taylor series expansion around $x=0$ up to the third term, the vertical force, F_{mv} can be represented as.

$$F_{mv} \cong \frac{F_m}{d} x - \frac{F_m}{2d^3} x^3 \quad (9)$$

It is noted that the vertical force, F_{mv} is a nonlinear third order polynomial function which matches the assumption earlier for magnetic response cantilever system. From the derived equations, it is observed that as the distance, d increases, the vertical force, F_{mv} will decrease accordingly. Since the repulsive force, F_m is also a function of distance, d , this simply indicates that as the relative distance between the magnets decreases, the resultant magnetic force applied on the cantilever system increases and hence the potential energy of the energy harvesting system will change accordingly.

As the mass, m being repositioned from its equilibrium point, the total force, F_{total} acting on the mass, m as shown in

Fig.3 will be the summation of the spring restoring force, F_s and the counter-restoring vertical force, F_{mv} that is described as.

$$F_{total} = F_s + F_{mv} = -k_{beam}x + \frac{F_m}{d}x - \frac{F_m}{2d^3}x^3 \quad (10)$$

The potential energy, $E(x)$ of the energy harvesting system is then deduced as.

$$E(x) = - \int F_{total} dx = \frac{1}{2} \left(k_{beam} - \frac{F_m}{d} \right) x^2 + \frac{F_m}{8d^3} x^4 \quad (11)$$

As the distance between the magnets increases, it is perceived that the repulsive magnetic force, F_m will also be reduced significantly and when it decreases until zero, the energy harvesting system will become a linear system. Fig. 4 shows the graph for potential energy, $E(x)$ of the energy harvesting system for diverse values of repulsive magnetic force per distance.

When the repulsive magnetic force per distance is less than or equal to the cantilever beam stiffness, k_{beam} , the potential energy, $E(x)$ of the energy harvester will have only one stable equilibrium position at its origin. In this circumstance, the system is said to be working under monostable condition and is still characterized under linear system. As the repulsive magnetic force per distance increases and exceeds the cantilever beam stiffness, k_{beam} , the potential energy, $E(x)$ will exhibit a bistable behavior whereby a symmetric double well with an unstable equilibrium position at the origin and two stable equilibrium positions exist within the energy harvester. When the system is working under bistable condition, the cantilever beam will oscillate within the stable equilibrium positions in each energy well as it will interchange from one equilibrium state to another when the external vibration energy is sufficiently high.

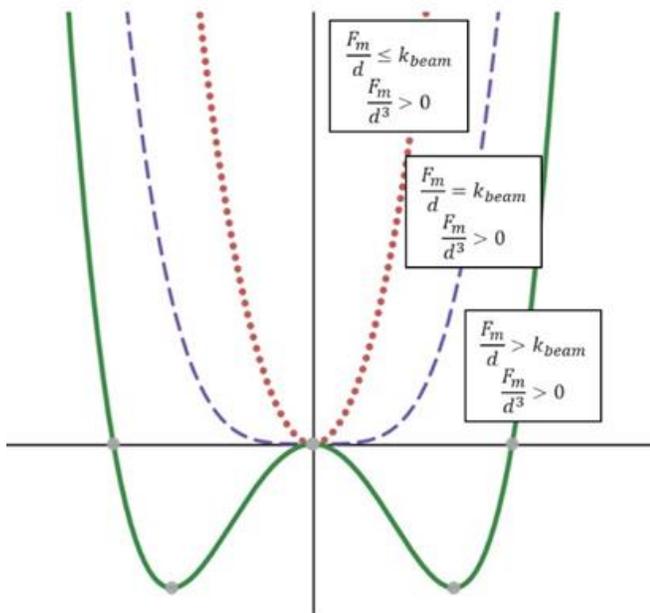


Fig. 4. Potential Energy, $E(x)$ for Monostable and Bistable Conditions.

The change in the potential energy of the cantilever beam basically results in a monostable or bistable characteristics in the energy harvesting system allowing the system to work differently based on the diverse requirements in the vibration environment. If the system is working under bistable behavior, the cantilever beam is able to switch between two equilibrium states, thus expanding the possibility of harvesting energy in a wider spectrum of vibrational frequency. Therefore, in order to achieve bistability in the system, the distance between the magnets should be sufficiently low enough to achieve the condition where the repulsive magnetic force per distance exceeds the stiffness of the cantilever beam, k_{beam} .

C. Effect of Magnetic Stiffness on Resonance Frequency

For a linear piezoelectric harvester with effective mass, m , the resonant frequency of the linear energy harvesting system can be described as.

$$\omega_{beam} = \sqrt{\frac{k_{beam}}{m}} \quad (12)$$

When the tip mass is replaced with a magnet and the energy harvesting system is deliberated as presented in Fig. 1, the resultant magnetic force will vary the stiffness of the cantilever beam, k_{beam} resulting in a change of resonance frequency. Consequently, the resonance frequency of the energy harvester will now be dependent on the stiffness linked to the magnetic force as well as the rigidity of the cantilever beam. The effective stiffness and the weight of the magnet will essentially determine the resonant frequency of the energy harvesting system. In order to parametrically model the influence of the magnetic strength and spectrum on the stiffness of the cantilever beam, a lumped parameter model was proposed to characterize the nonlinear system.

Fig. 5 illustrates the lumped parameter model of the nonlinear energy harvesting system as a consequence of additional stiffness acquainted with the repulsive magnetic force. The resultant force contributed by the repulsive magnetic force is demonstrated as a variable spring where the stiffness of the spring is subjected to the variation of the magnetic force with respect to the relative distance between both of the magnets.

As a result, the total effective stiffness and its resonant frequency of the energy harvester can be further deduced as

$$K_{eff} = K_{beam} + K_{magnet} \quad (13)$$

$$\omega_{eff} = \sqrt{\frac{K_{eff}}{m_{eff}}} \quad (14)$$

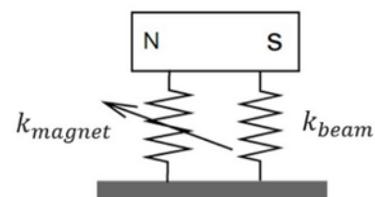


Fig. 5. Lumped Parameter Model for Nonlinear Energy Harvesting System.

Theoretically, as the magnetic force on the beam increases, the stress/strain imposed on the cantilever beam will also increase accordingly. Consequently, the resonance frequency will also change depending on the attractive or repulsive mode of the magnets and the resultant stiffness at the mass tip of the cantilever. It should be noted that stress induced by the magnets should not be larger than the yield stress of the beam for repulsive magnetic mode. By deducing the equations, one can tune the desired resonant frequency, ω_{tuned} from the original beam frequency, ω_{beam} of the energy harvester by determining the appropriate magnetic stiffness using the equation.

$$K_{magnet} = m_{eff}\omega_{tuned}^2 - K_{beam} = m_{eff}(\omega_{tuned}^2 - \omega_{beam}^2) \quad (15)$$

III. RESULT AND DISCUSSION

A. Design of Experiment

To assess the feasibility of the nonlinear energy harvester, a prototype of the system was built and verified in the laboratory. Fig. 6 shows the apparatus and design of the experimental structure for the energy harvesting system. The vibration shaker was used to produce the controllable vibration base excitation to the whole system throughout the experiment. A piezoelectric strip with dimensions of 1.25 x 0.5 x 0.02 inch, weighing 1.372 gram was used as the cantilever beam in the system. Fig. 7 illustrates the dimension of the piezoelectric cantilever applied throughout the experiment. A permanent magnet weighing 0.75 gram was secured at the free end tip of the cantilever beam while another magnet with opposing polarity was fixed on an external stationary structure with the purpose to adjust the distance between the magnets. The magnets were arranged in repulsive mode and were displaced in the direction of x -axis.

The empirical study of the bistable energy harvesting system was divided into two classifications. The initial part of the experiment comprises of analysis on the effect of the spacing displacement, d between two of the repulsive magnets on the resonant frequency of the system. In the subsequent experiment, the performance in terms of voltage output of the piezoelectric energy harvester was further examined according to the variety of the relative distance between the two repulsive magnets. Through the variation of distance between the magnetic mass, the stiffness of the cantilever beam can be regulated to obtain the desired frequency range. As the relative displacement between both of the repulsive magnets decreases, the stationary location of the cantilever beam changes accordingly. The variation of stationary point is reliant on the extent of hardening influence on the piezoelectric cantilever as the relative displacement between the two repulsive magnets changes.

Furthermore, as discussed in the previous sections, both of these experiments were done under two different circumstances scenarios. One of the scenarios involves pairing of the static magnetic mass under stationary condition while another scenario involves pairing of the static magnetic mass under similar vibration source, as depicted in Fig. 1. The frequency response curve for both scenarios was investigated respectively. The impact of spacing displacement on the magnetic force that affects the bistability of the energy harvesting system was also observed accordingly.

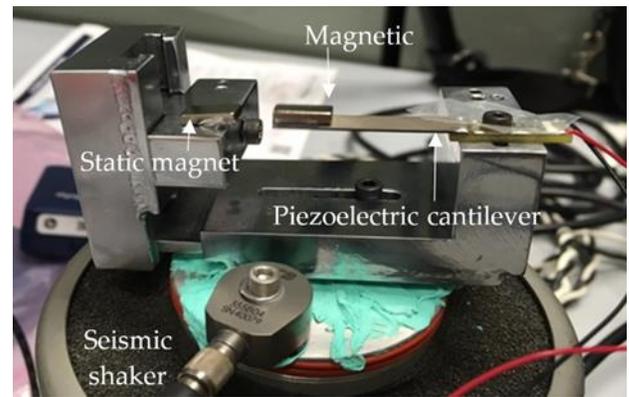


Fig. 6. Design of Experiment Apparatus of the Piezoelectric Energy Harvester.

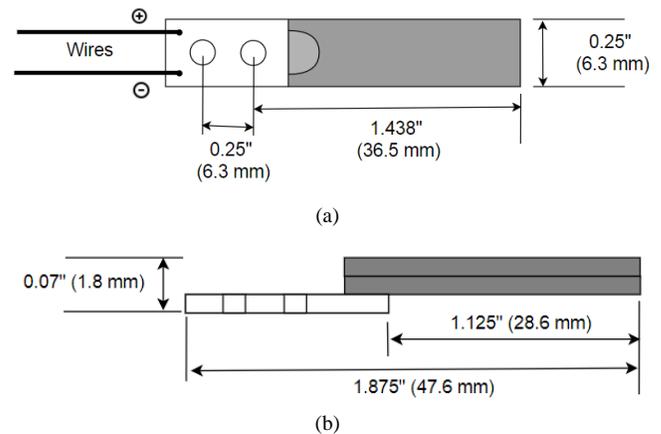


Fig. 7. Dimension of the Piezoelectric Cantilever Beam (a) Top View (b) Side View.

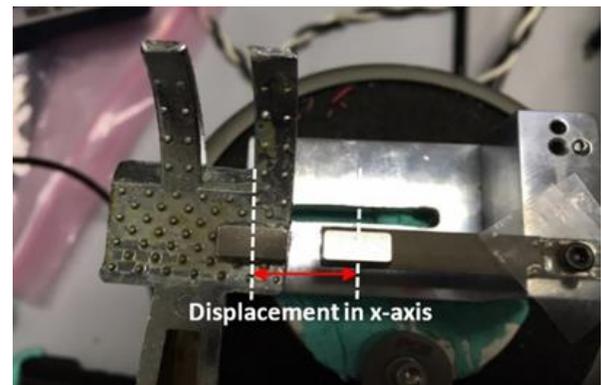


Fig. 8. Top View of Experiment Apparatus of the Piezoelectric Energy Harvester with Displacement in x -axis.

Fig. 8 illustrates the top view of the apparatus setup with horizontal spacing displacement in x -axis between the two magnets. For this setup, the energy harvester involves pairing of the static magnetic mass under stationary condition. During the experiment, the static magnet was adjusted and move away from the piezoelectric cantilever beam to observe the impact of magnetic force on the system. The spacing displacement between the magnets was measured from the center of the respective repulsive magnets.

B. Effect of Spacing Displacement on Resonant Frequency

An analysis on the effect of applying magnetic forces in the adjustment of the resonance frequency of the energy harvesting system has been investigated as follows. In the first part of the experiment, the fixed magnet is placed under stationary condition to observe the influence on the resonant frequency of the energy harvester as presented in Fig. 1(a). The static magnet is then displaced along the x -axis, making sure there is no displacement in the y -axis and z -axis between the two repulsive magnets in the system. The deviations of the resonant frequency as a function of spacing distance for PSM system is then plotted as presented in Fig. 9(a). Based on the observation, as the magnets were moved closer to each other from large distances, the system tends to behave as monostable linear system with decreasing resonance frequency until it reaches the condition where the repulsive magnetic force per distance, exceeds the stiffness of the cantilever beam, k_{beam} . Subsequent to this condition, when the magnets are pushed further and are sufficiently close, the resonant frequency tend to increase drastically creating bistability condition for the system. In this circumstance, the system has the ability to interchange between two equilibrium states and the fluctuation of the tip displacement can surge substantially if the base excitation level is increased sufficiently.

Further analysis on the PSM system results shows that for spacing displacement above 8 mm as in region IV, the resonant frequency of the energy harvesting system tends to stabilize, indicating the waning of repulsive force by the magnets. As the relative displacement between the magnets decreases, the resonant frequency of the system seems to decrease accordingly, as shown in region III. It is also observed that for displacement ranges between 3 mm and 4 mm (region II), there is a sharp decrease of resonant frequency up to its minimum frequency of 114 Hz. It is then followed by a sharp increase of resonant frequency (region I) as we decrease the spacing displacement, placing the magnets into stronger repulsive mode. As the strength of the repulsive magnet force increases, the piezoelectric cantilever beam responded accordingly through the change in its static position and bending effect, subsequently triggering a drastic adjustment in the resonant frequency response. At this point, the energy harvester will be working under bistable condition. Thus, the resultant analysis displays that in order to ensure that the energy harvesting system is working under bistable mode; the minimum requirement for the distance between the repulsive magnets is less than 3 mm.

In the second part of the experiment, the fixed magnet is placed under the same vibration source as the piezoelectric cantilever beam to observe the effect on its resonant frequency as demonstrated in Fig. 1(b). Fig. 9(b) illustrates the variations of the resonant frequency when the spacing distance between two magnets is altered accordingly under the common vibration source for PVM system. It is observed that a similar pattern of resonant frequency is obtained in comparison with the PSM system. However, the change of resonant frequency may not be as abrupt compared to the static magnetic mass

under stationary condition. Based on the analysis, the resonant frequency of the system stabilizes as the distance between both magnets increases above 6mm, as shown in region III. As we move the magnets to closer to each other, the resonant frequency of the PVM system decreases gradually for distances between 4 mm to 6 mm (region II). In this setup, the energy harvesting system began to function as a bistable system for magnetic distances below 4 mm (Region I).

For the significance of this investigation, it is perceived that the repulsive influence between the magnets tend to stabilize only at displacement beyond 6 mm for the energy harvesting system. For relatively smaller distance of the spacing displacement, the increase of hardening influence between the magnets causes further bending effect on the cantilever and thus creates bistability of the resonant frequency in the system. In comparison between both of the systems, the PVM system seems to provide more feasible design for the energy harvester and it was able to achieve its bistability mode for larger threshold distance between magnets. This proves that the design of the nonlinear energy harvester should include the effect of vibration source on both the piezoelectric cantilever beam as well as the fixed magnet on the external structure.

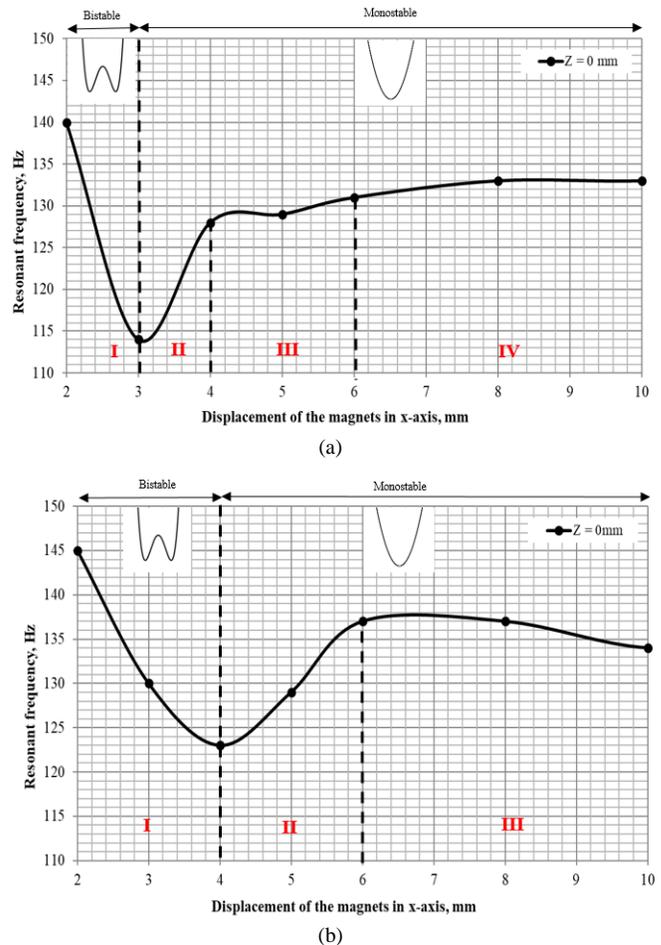


Fig. 9. Resonant Frequency of the Energy Harvester for different the Spacing Displacement, d for Static Magnetic Mass under (a) PSM (b) PVM.

Furthermore, the resultant graph for both experiments are also consistent with the theoretical expectations, endorsing that the cantilever beam leaps between two stable states for sufficiently close magnetic distance and adequate base excitation as demonstrated in Fig. 4. This change in resonance frequency due to the spacing displacement of the magnets was also found to have similar pattern as other researchers' work [43, 44]. However, in comparison with the other researchers' work, it was found that the addition of vibration source on the external magnet increases the region of bistability for the energy harvester. The resonance frequency of piezoelectric cantilever for the energy harvester with same vibration source also increases in comparison with the stationary magnet due to the impact of vibration on the coupling between the magnets.

These findings are substantially vital during the design considerations for vibration energy harvester and notable for researchers in this field of research.

C. Effect of Spacing Displacement on Output Voltage

A similar setup as presented in Fig. 1 was applied to measure and verify the performance of the nonlinear energy harvester in terms of open-circuit output voltage. The energy harvester was excited by a controllable vibration shaker with mechanical vibrations and placed accordingly using both scenarios as discussed in Section II. The excitation level of the shaker is set sufficiently low to prevent initiation of bistable mode as the distance between the magnets becomes substantially close. Consequently, the energy harvester was held to oscillate at only one interwell equilibrium state throughout the experiment. The function generator and amplifier was used to regulate the vibration shaker in tuning the resonance frequency for each cantilever beams. The output electrical responses were observed using the oscilloscope and its frequency response curve are plotted accordingly.

The resultant graph will also include the voltage output of the linear system with proof mass for comparison purposes. Fig. 10 illustrates the frequency response output of the open-circuit root mean square (rms) voltage generated by the energy harvester for a variation of relative displacement between the magnets for the PSM and PVM system. The differences in the output voltage were compared to observe the impact of magnetic force and location of vibration source on the cantilever beam. Based on our observation, the results showed that the relative displacement between the two magnets significantly affects the level of the harvested voltage. It is also eminent that the positioning of the vibration source on the piezoelectric cantilever and/or permanent magnets impacts the frequency response curves especially in terms broadening the frequency range of the energy harvester.

Fig. 10 shows that the relative distance between the two magnets significantly affects the level of the harvested voltage. It is also observed that by changing the relative displacement, the energy harvesting system has the competence to harvest vibrational energy in a wider spectrum of operating frequency. For the spacing displacement of $d=2\text{mm}$ as presented in Fig. 10(a) and Fig. 10(b), since the magnets were placed essentially close to each other, both of the PSM and PVM systems are now under bistable conditions. At this point, the magnetic field strength applied on the system was very high and the nonlinear

characteristics will be substantially visible as the excitation level of the vibration source increases. However, the effect of magnet on the piezoelectric cantilever causes the output voltage to drop significantly lower than its other counterparts. The resultant bandwidth for both of the system is also lower and does not provide a good model for broadband energy harvesting.

For the spacing displacement of $d=3\text{mm}$ as shown in Fig. 10(a), the PSM system was operating at its lowest resonance frequency as discussed in the previous section. In this case, both PSM and PVM systems are still operating in bistable conditions. However, the PSM system tends to exhibit broader frequency range as compared to the linear system and its other counterparts. Similarly, at the spacing displacement of $d=4\text{mm}$ for PVM system as shown in Fig. 10(b), where PVM was also operating at its lowest resonance frequency, the output voltage and its bandwidth are significantly higher than its counterparts of different spacing distances. In other words, in designing a bistable system, the optimal spacing distance to be chosen is when the resonant frequency of the cantilever beam is at its lowest point. The magnets were displaced further at spacing distance of 5 mm and above. The frequency response curves in Fig. 10(a) show that the output voltage of the PSM system decreases and remain fluctuating around the similar output voltage accordingly. Similarly, at Fig. 10(b), the output voltage remains constantly the same for spacing distance of 6mm and above.

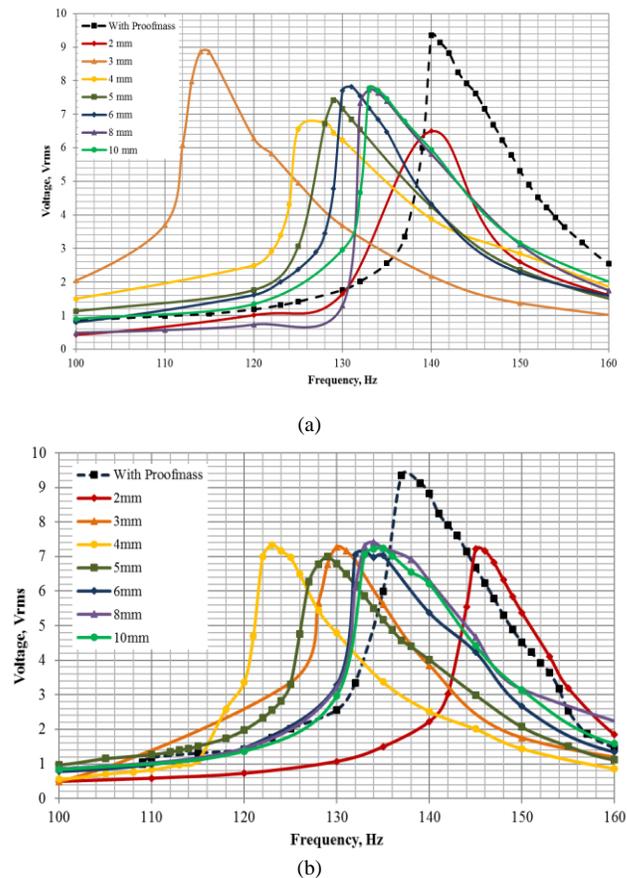


Fig. 10. Output Voltage of the Energy Harvester for different the Spacing Displacement, d for Static Magnetic Mass under (a) PSM (b) PVM.

These outputs indicate that both the PSM and PVM system were not affected significantly by the repulsive force of the magnets and thus regardless of the position of the vibration source, both system exhibits similar frequency response curve. However, it is noticeable that the output voltage for both PSM and PVM system is slightly lower in comparison with the linear system with proof mass and yet it has similar spectrum of frequency range in the frequency response curves.

As can be perceived from the experimental figures, with the decrease of spacing distance between the magnets, the impact of magnetic force intensifies and thus increases the hardening response of the piezoelectric cantilever. This notion of experiment indicated the ability of the energy harvesting system to shift from linear to nonlinear system gradually using permanent magnets. The results also showed that the position of vibration source affects the resonance frequency as well as the operating frequency of the energy harvesting system. Note that as the spacing distance was decreased to its minimal length, the frequency response did not provide the best bandwidth for PVM system. This indicates that as the magnetic force increases, the increased hardening response in the piezoelectric cantilever can cause reduction in bandwidth and eventually may not provide the widest possible range of operating frequency to harvest the vibrational energy. Thus, it is essential to note that the relative distance between two repulsive magnets as well as the positioning of vibration source can directly affect the resonant frequency and the operational range of frequency spectrum of the energy harvester.

IV. CONCLUSION

In this paper, the concept of bistable piezoelectric energy harvesting system by exploiting the effect of magnetic force was presented. Detailed investigation on the effect of the horizontal spacing distance between two magnets of repulsive mode in a piezoelectric energy harvester has also been deliberated. The impact of positioning the vibration source on either the piezoelectric cantilever beam only or both the piezoelectric cantilever beam and external magnets was compared in terms of resonance frequency and bandwidth of the operating frequency for broadband energy harvesting. It is shown that the model is feasible and applicable in harvesting energy in a wider band of frequency spectrum from ambient mechanical vibrations. Likewise, it is also established that the fundamental resonant frequency of the cantilever beam has significant dependence on the variations of relative distance between the magnets as well as the positioning of vibration source on the external magnets.

This paper shows that the piezoelectric energy harvester can be envisioned to accomplish wideband frequency energy harvesting by selecting the optimal spacing distance between two repulsive magnets. Alteration of spacing displacement can also be used to adjust the resonant frequency of the harvester for matching of surroundings excitation frequency in the environment.

V. FUTURE WORK

As a continuation of this research, further work on the impact of different sizes and locations of the magnets on the cantilever beam can be considered. The size of the magnets

may impart a different magnetic force towards the stress and strain of the cantilever beams and is therefore noteworthy for future work considerations. Furthermore, multiple piezoelectric beams with magnets and different configurations can also be worked upon for further considerations in increasing the bandwidth of the vibration energy harvesting system.

ACKNOWLEDGMENT

The authors would like to express gratitude and sincere appreciation to the Ministry of Higher Education of Malaysia for the financial support through Fundamental Research Grant Scheme (FRGS/1/2020/TK0/MMU/03/13). The authors would also like to acknowledge the Faculty of Engineering and Technology, Multimedia University and Faculty of Electronic and Computer Engineering, Universiti Teknikal Malaysia Melaka for the support given in conducting this research.

REFERENCES

- [1] J. Kim, A. S. Campbell, B. E. F. de Ávila, and J. Wang, "Wearable biosensors for healthcare monitoring," *Nat. Biotechnol.*, vol. 37, no. 4, pp. 389–406, April 2019. <https://doi.org/10.1038/s41587-019-0045-y>.
- [2] A. E. Akin-Ponnle and N. B. Carvalho, "Energy Harvesting Mechanisms in a Smart City—A Review," *Smart Cities*, vol. 4, no. 2, pp. 476–498, April 2021. <https://doi.org/10.3390/smartcities4020025>.
- [3] J. Curry and N. Harris, "Powering the environmental Internet of Things," *Sensors*, vol. 19, no. 8, pp.1940, April 2019. <https://doi.org/10.3390/s19081940>.
- [4] A.A.A. Zayed, S.F.M. Assal, K. Nakano, T. Kaizuka, and A.M.R. Fath El-Bab, "Design Procedure and Experimental Verification of a Broadband Quad-Stable 2-DOF Vibration Energy Harvester," *Sensors*, vol. 19, pp. 2893, 2019. <https://doi.org/10.3390/s19132893>.
- [5] A. Kumar, S.F. Ali, and A. Arockiarajan, "Exploring the benefits of an asymmetric monostable potential function in broadband vibration energy harvesting," *Appl. Phys. Lett.*, vol. 112, pp. 233901, 2018. <https://doi.org/10.1063/1.5037733>.
- [6] M.G. Kang, W.S. Jung, C.Y. Kang, and S.J. Yoon, "Recent Progress on PZT Based Piezoelectric Energy Harvesting Technologies," *Actuators*, vol. 5, no. 1, pp. 5, 2016. <https://doi.org/10.3390/act5010005>.
- [7] K. Li, X. He, X. Wang, and S. Jiang, "A Nonlinear Electromagnetic Energy Harvesting System for Self-Powered Wireless Sensor Nodes," *J. Sens. Actuator Netw.*, vol. 8, pp. 18, 2019. <https://doi.org/10.3390/jsan8010018>.
- [8] S. Mohammadi and A. Esfandiari, "Magnetostrictive vibration energy harvesting using strain energy method," *Energy*, vol. 81, pp. 519–525, 2015. <https://doi.org/10.1016/j.energy.2014.12.065>.
- [9] S. Priya, H. Song, Y. Zhou, R. Varghese, A. Chopra, S. Kim, I. Kanno, L. Wu, D. Ha, J. Ryu, and R. Polcawich, "A Review on Piezoelectric Energy Harvesting: Materials, Methods, and Circuits," *Energy Harvest. Syst.*, vol. 4, pp. 3–39, 2017. <https://doi.org/10.1515/ehs-2016-0028>.
- [10] G. Scarselli, F. Nicassio, F. Pinto, F. Ciampa, O. Iervolino, and M. Meo, "A novel bistable energy harvesting concept," *Smart Mater Struct.*, vol. 25, pp. 055001, 2016. <https://doi.org/10.1088/0964-1726/25/5/055001>.
- [11] S.E. Jo, M.S. Kim, and K.J. Kim, "A resonant frequency switching scheme of a cantilever based on polyvinylidene fluoride for vibration energy harvesting," *Smart Mater Struct.*, vol. 21, pp. 015007, 2011. <https://doi.org/10.1088/0964-1726/21/1/015007>.
- [12] A. H. Alameh, M. Gratuze, and F. Nabki, "Impact of geometry on the performance of cantilever-based piezoelectric vibration energy harvesters," *IEEE Sens. J.*, vol. 19, pp. 10316–10326, 2019. <https://doi.org/10.1109/JSEN.2019.2932341>.
- [13] A. Hajati and S.G. Kim, "Ultra-wide bandwidth piezoelectric energy harvesting," *Appl. Phys. Lett.*, vol. 99, pp. 083105, 2011. <https://doi.org/10.1063/1.3629551>.
- [14] S. Zhou, W. Chen, M.H. Malakooti, J. Cao, and D.J. Inman, "Design and modeling of a flexible longitudinal zigzag structure for enhanced

- vibration energy harvesting,” *J Intell Mater Syst Struct.*, vol. 28, pp. 367-380, 2017. <https://doi.org/10.1177/1045389X16645862>.
- [15] S. Kumar, A. Mitra, and H. Roy, “Geometrically nonlinear free vibration analysis of axially functionally graded taper beams,” *Int. J. Eng. Sci. Technol.*, vol. 18, pp. 579-593, 2015. <https://doi.org/10.1016/j.jestch.2015.04.003>.
- [16] Y. Uzun and E. Kurt, “The effect of periodic magnetic force on a piezoelectric energy harvester,” *Sens. Actuator A Phys.*, vol. 192, pp. 58-68, 2013. <https://doi.org/10.1016/j.sna.2012.12.017>.
- [17] J. Jiang, S. Liu, L. Feng, and D. Zhao, “A review of piezoelectric vibration energy harvesting with magnetic coupling based on different structural characteristics,” *Micromachines*, vol. 12, pp. 436, 2021. <https://doi.org/10.3390/mi12040436>.
- [18] J. Liang, “Synchronized bias-flip interface circuits for piezoelectric energy harvesting enhancement: A general model and prospects,” *J Intell Mater Syst Struct.*, vol. 28, pp. 339-356, 2017. <https://doi.org/10.1177/1045389X16642535>.
- [19] E. Lefeuvre, A. Badel, A. Brenes, S. Seok, and C.S. Yoo, “Power and frequency bandwidth improvement of piezoelectric energy harvesting devices using phase-shifted synchronous electric charge extraction interface circuit,” *J Intell Mater Syst Struct.*, vol. 28, pp. 2988-2995, 2017. <https://doi.org/10.1177/1045389X17704914>.
- [20] M.R. Mhetre and H.K. Abhyankar, “Human exhaled air energy harvesting with specific reference to PVDF film,” *Int. J. Eng. Sci. Technol.*, vol. 20, pp. 332-339, 2017. <https://doi.org/10.1016/j.jestch.2016.06.012>.
- [21] Z. Lin, J. Yang, J. Zhao, N. Zhao, J. Liu, Y. Wen, and P. Li, “Enhanced broadband vibration energy harvesting using a multimodal nonlinear magnetoelectric converter,” *J. Electron. Mater.*, vol. 45, pp. 3554–3561, 2016. <https://doi.org/10.1007/s11664-016-4531-4>.
- [22] Y. Cheng, N. Wu, and Q. Wang, “An efficient piezoelectric energy harvester with frequency self-tuning,” *J. Sound Vib.*, vol. 396, pp. 69-82, 2017. <https://doi.org/10.1016/j.jsv.2017.02.036>.
- [23] Y.-H. Shin, J. Choi, S.J. Kim, S. Kim, D. Maurya, T.-H. Sung, S. Priya, C.-Y. Kang, and H.-C. Song, “Automatic resonance tuning mechanism for ultra-wide bandwidth mechanical energy harvesting,” *Nano Energy*, vol. 77, pp. 104986, 2020. <https://doi.org/10.1016/j.nanoen.2020.104986>.
- [24] M. Lallart, S.R. Anton, and D.J. Inman, “Frequency self-tuning scheme for broadband vibration energy harvesting,” *J Intell Mater Syst Struct.*, vol. 21, pp. 897-906, 2010. <https://doi.org/10.1177/1045389X10369716>.
- [25] L. Yu, L. Tang, and T. Yang, “Piezoelectric passive self-tuning energy harvester based on a beam-slider structure,” *J. Sound Vib.*, vol. 489, pp. 115689, 2020. <https://doi.org/10.1016/j.jsv.2020.115689>.
- [26] W. Yang and S. Towfighian, “A parametric resonator with low threshold excitation for vibration energy harvesting,” *J. Sound Vib.*, vol. 446, pp. 129-143, 2019. <https://doi.org/10.1016/j.jsv.2019.01.038>.
- [27] M. Lallart, C. Richard, L. Garbuio, L. Petit, and D. Guyomar, “High efficiency, wide load bandwidth piezoelectric energy scavenging by a hybrid nonlinear approach,” *Sens. Actuator A Phys.*, vol. 165, pp. 294-302, 2011. <https://doi.org/10.1016/j.sna.2010.09.022>.
- [28] Z. Yi, Y. Hu, B. Ji, J. Liu, and B. Yang, “Broad bandwidth piezoelectric energy harvester by a flexible buckled bridge,” *Appl. Phys. Lett.*, vol. 113, pp. 183901, 2018. <https://doi.org/10.1063/1.5049852>.
- [29] D. Gibus, P. Gasnier, A. Morel, F. Formosa, L. Charleux, S. Boisseau, G. Pillonnet, C.A. Berlitz, A. Quelen, and A. Badel, “Strongly coupled piezoelectric cantilevers for broadband vibration energy harvesting,” *Appl. Energy*, vol. 277, pp. 115518, 2020. <https://doi.org/10.1016/j.apenergy.2020.115518>.
- [30] R.M. Toyabur, M. Salauddin, and J.Y. Park, “Design and experiment of piezoelectric multimodal energy harvester for low frequency vibration,” *Ceram. Int.*, vol. 43, pp. S675-S681, 2017. <https://doi.org/10.1016/j.ceramint.2017.05.257>.
- [31] E. Dechant, F. Fedulov, D.V. Chashin, L.Y. Fetisov, Y.K. Fetisov, and M. Shamonin, “Low-frequency, broadband vibration energy harvester using coupled oscillators and frequency up-conversion by mechanical stoppers,” *Smart Mater Struct.*, vol. 26, pp. 065021, 2017. <https://doi.org/10.1088/1361-665X/aa6e92>.
- [32] P. Li, S. Gao, X. Zhou, H. Liu, and J. Shi, “Analytical modeling, simulation and experimental study for nonlinear hybrid piezoelectric---electromagnetic energy harvesting from stochastic excitation,” *Microsyst. Technol.*, vol. 23, pp. 5281–5292, 2017. <https://doi.org/10.1007/s00542-017-3329-5>.
- [33] C. Lan and W. Qin, “Enhancing ability of harvesting energy from random vibration by decreasing the potential barrier of bistable harvester,” *Mech Syst Signal Process.*, vol. 85, pp. 71-81, 2017. <https://doi.org/10.1016/j.ymsp.2016.07.047>.
- [34] W. Wang, J. Cao, C.R. Bowen, D.J. Inman, and J. Lin, “Performance enhancement of nonlinear asymmetric bistable energy harvesting from harmonic, random and human motion excitations,” *Appl. Phys. Lett.*, vol. 112, pp. 213903, 2018. <https://doi.org/10.1063/1.5027555>.
- [35] H. Xingbao and B. Yang, “Improving energy harvesting from impulsive excitations by a nonlinear tunable bistable energy harvester,” *Mechanical Systems and Signal Processing*, vol. 158, pp. 107797, 2021. <https://doi.org/10.1016/j.ymsp.2021.107797>.
- [36] P. Deepak and B. George, “Piezoelectric energy harvesting from a magnetically coupled vibrational source,” *IEEE Sens. J.*, vol. 21, pp. 3831-3838, 2021. <https://doi.org/10.1109/JSEN.2020.3025216>.
- [37] Z. Ren, H. Zhao, C. Liu, L. Qian, S. Zhang, and J. Zhao, “Study the influence of magnetic force on nonlinear energy harvesting performance,” *AIP Advances*, vol. 9, pp. 105107, 2019. <https://doi.org/10.1063/1.5111848>.
- [38] P. Firoozy, S.E. Khadem, and S.M. Pourkiaee, “Broadband energy harvesting using nonlinear vibrations of a magnetopiezoelectric cantilever beam,” *Int. J. Eng. Sci.*, vol. 111, pp. 113–133, 2017. <https://doi.org/10.1016/j.ijengsci.2016.11.006>.
- [39] N. Jackson, “PiezoMEMS nonlinear low acceleration energy harvester with an embedded permanent magnet,” *Micromachines*, vol. 11, pp. 500, 2020. <https://doi.org/10.3390/mi11050500>.
- [40] M. Kim, M. Hoegen, J. Dugundji, and B.L. Wardle, “Modeling and experimental verification of proof mass effects on vibration energy harvester performance,” *Smart Mater. Struct.*, vol. 19, pp. 045023, 2010. <https://doi.org/10.1088/0964-1726/19/4/045023>.
- [41] A. Erturk and D.J. Inman, “An experimentally validated bimorph cantilever model for piezoelectric energy harvesting from base excitations,” *Smart Mater. Struct.*, vol. 18, pp. 025009, 2009. <https://doi.org/10.1088/0964-1726/18/2/025009>.
- [42] S. Zhou, J. Cao, and J. Lin, “Theoretical analysis and experimental verification for improving energy harvesting performance of nonlinear monostable energy harvesters,” *Nonlinear Dyn.*, vol. 86, pp. 1599–1611, 2016. <https://doi.org/10.1007/s11071-016-2979-7>.
- [43] M. Ferrari, V. Ferrari, M. Guizzetti, B. Andò, S. Baglio, and C. Trigona, “Improved energy harvesting from wideband vibrations by nonlinear piezoelectric converters,” *Sens. Actuator A Phys.*, vol. 162, pp. 425-431, 2010. <https://doi.org/10.1016/j.sna.2010.05.022>.
- [44] V. Shah, R. Kumar, M. Talha, and J. Twiefel, “Numerical and experimental study of bistable piezoelectric energy harvester,” *Integr. Ferroelectr.*, vol. 192, pp. 38-56, 2018. <https://doi.org/10.1080/10584587.2018.1521669>.

Smart Blended Learning Framework based on Artificial Intelligence using MobileNet Single Shot Detector and Centroid Tracking Algorithm

Abdul Wahid¹, Muhammad Fajar B², Jumadi M. Parenreng³, Seny Luhriyani⁴, Puput Dani Prasetyo Adi⁵
Computer Engineering Study Program, Universitas Negeri Makassar, Makassar, Indonesia^{1,2,3}
English Department, Universitas Negeri Makassar, Makassar, Indonesia⁴
National Research and Innovation Agency (BRIN), Bandung, Indonesia⁵

Abstract—The Covid-19 pandemic has affected all aspects of human life and has even forced humans to shift their life habits, including in the world of education. The learning model must shift from the traditional face-to-face pattern to a modern face-to-face pattern or an asynchronous pattern with information technology-based applications. Blended learning is one of the appropriate solutions to adjust the limited face-to-face learning conditions. Blended learning can be done, for example, by scheduling learning by dividing the number of participants by 50% and entering on a scheduled basis. However, the problem is that the time and effort used are less efficient. Blended learning can also be done by conducting learning simultaneously with 50% of students in class and the remaining 50% through conferences. This concept will streamline the time and effort used. However, the problem is that there is a gap in the learning experience between students in class and students who do learning via conference. This innovative blended learning system framework is proposed to overcome these problems. The system built seeks to present an online learning experience atmosphere so that it is expected to be able to resemble an offline learning atmosphere. We created a system using camera technology and object detection that will track the movement of the teacher so that the teacher can move freely in the room without having to be stuck in front of the computer holding the conference. The algorithms used are MobileNet Single Shot Detector and Centroid Tracking. This research produces an accurate model for detecting teacher movement at a distance of 2, 4, and 6 meters with a camera installation height of 1.5 and 3 meters.

Keywords—Smart blended learning; mobilenet; single shot detector; convolutional neural network; centroid tracking

I. INTRODUCTION

Blended learning has been applied in higher education for several years, but there is still limited research on what affects student satisfaction in blended learning environments in higher education [1]. The implementation of blended learning at a university is evaluated by measuring student satisfaction in face-to-face sessions, independent study sessions using online learning, and the overall learning experience in a Blended Learning environment.

There are two main areas related to the blended learning environment. The first is a blend of traditional classroom learning and e-learning, and the second is synchronous and asynchronous e-learning technology [2]. This first area is the

best-known form of combination seen in the amalgamation of theory and practice of instructor and student-centered learning. The second type of blended learning is the blend of technologies that give students access to synchronous and asynchronous communication and information. This is especially useful when considering the number of external students outside of campus studying at the tertiary level and the associated geographic and access issues and creating an environment accommodating cross-cultural learners.

Along with the development of artificial intelligence technology, various technologies are born to assist humans in completing a task or activity. The technology in question is capable of performing actions like a human. An example is computer intelligence replacing the human sense of sight, usually known as computer vision [3]. By utilizing camera functions supported by object detection algorithms, today's computers can intelligently carry out surveillance automatically like a human's vision.

On the one hand, the current Covid-19 pandemic requires physical interaction restrictions to prevent virus transmission [4]. These restrictions have brought significant changes in various fields, especially in education, namely the limited offline learning activities that are shifted to semi-online and even full-online learning to avoid interactions that can spread virus transmission.

One of the benefits of blended learning is that learning becomes more flexible because its implementation is not limited by distance and time. Online learning from home and offline learning in the classroom can be carried out simultaneously [5]–[9]. However, there are shortcomings, namely the limited space for teachers and students in learning activities. There is a gap between the learning experience in offline classes and online learning conditions. Therefore, a system can be developed to overcome this gap by utilizing computer vision technology, especially object detection technology. The system in question is a system that can use a camera to detect objects on the teacher so that the teacher has free space to explain like in an offline class, and the teacher's position will still focus on video because of the combination of objects detection algorithms and tracking algorithms. The benefit for students is to present a learning experience that tries to approach conditions like in an offline class.

II. LITERATURE STUDY

Object tracking is essential in computer vision and widely used in human-computer interaction, surveillance, and medical imaging. In its simplest form, tracking can be defined as estimating the trajectory of an object in the image plane as it moves around the frame [10]. Object tracking has attracted significant attention because it can perform a wide range of processes, including intelligence in video surveillance, machine and human interfaces, and the field of robotics [11]. However, designing an excellent visual tracking method is still an open issue. Challenges in visual tracking problems include varying shapes and appearances of objects, occlusion, lighting changes, irregular scenes, etc.

The object tracking process consists of two stages in analyzing the video, namely detecting objects and tracking the movement of objects from frame to frame [12]. One of the tracking algorithms that can be used is the centroid tracking algorithm. The centroid tracking algorithm works by taking into account the center point (centroid) of an object in tracking [13]. Therefore, the tracking process in the centroid tracking algorithm is very dependent on the accuracy of the object detection or identification algorithm. The tracking process will be challenging if the object detection algorithm has minimal accuracy and takes a long time. An object tracking task requires an object detection process that can work in real-time and has good accuracy to avoid decision errors.

Research on object identification using a neural network has been carried out by Girshick by proposing the R-CNN architecture [14]. The results obtained are 66% accuracy with 47 seconds per image detection time. The long detection time is due to the algorithm classifying as many as two thousand proposal regions for each image. Furthermore, Girshick optimized the study [14] and proposed the Fast R-CNN architecture [15]. This algorithm can shorten the detection time to 25 times faster than R-CNN and produces an accuracy of 66.9%. The increase in detection speed occurs because the CNN process was initially carried out amounted to about two thousand times be reduced to only one time. The detection time per image only takes 2 seconds.

Research by Ren et al. [16] proposed an architecture called Faster R-CNN. The idea of Faster R-CNN is to replace the selective search algorithm to generate proposal regions used in research [14], [15], becoming a Region Proposal Network (RPN). The detection time per image can reach 0.2 seconds with an accuracy of 66.9%.

Li et al. [17] carried out subsequent research to test the performance of an architecture called MobileNet SSD. The test results show that the accuracy can reach 95% with a detection time of 0.12 seconds. MobileNet SSD combines MobileNet architecture and Single Shot Multibox Detector (SSD) that uses depthwise and pointwise convolution, reducing computation significantly [18]. MobileNet SSD can provide faster object detection performance compared to Faster R-CNN.

Rahman et al. [13] conducted research on object detection in the form of vehicles to find out which vehicles are against

the current or in the opposite direction. The movement of objects is known to use the YOLO detection method and centroid tracking. Centroid tracking works well on single class objects, but if applied to multiclass objects; it allows the identity of object tracking to be exchanged. The results of this study can track vehicles against the direction of the current, which is tested on a 1280x720 video.

Distance learning is usually carried out via video conferencing. In general, video conferencing use technology assistance such as WebRTC [19]. This technology allows video, audio, or other data to be transferred to the student in real-time. In blended learning, the problem is that the video source that is usually used is placed in a static position in front of the teacher so that the teacher can only stand in front of the video source camera. In our proposed research, object detection and tracking technology are implemented with the aim that the teacher's camera can be placed far from the teacher to shoot a wide area in the classroom so that wherever the teacher moves, the video will still focus on the teacher in the center of the video.

III. METHODS

Data collection is the first step in the research process—image data is produced by splitting test and training data. The test data is organized around a single object class: humans. The previously provided test data is fed into the pre-processed data. The data is pre-processed to create a more optimum feature extraction. The next step is to extract the features to obtain a value utilized as input in the following stage. MobileNet SSD is the object detection algorithm. The outcomes of the architectural trials are analyzed, and the best accuracy is chosen for use as training for future test data. As for the test data, pre-processed data is carried out, and then we proceeded with feature extraction and final testing. Workflow can be seen in Fig. 1.

A. Data Acquisition Stage

The data used as input is a real-time video recording. Aside from video recordings, the dataset was derived from open-source internet sources, specifically Open Images and YouTube videos containing things to be detected, specifically human objects.

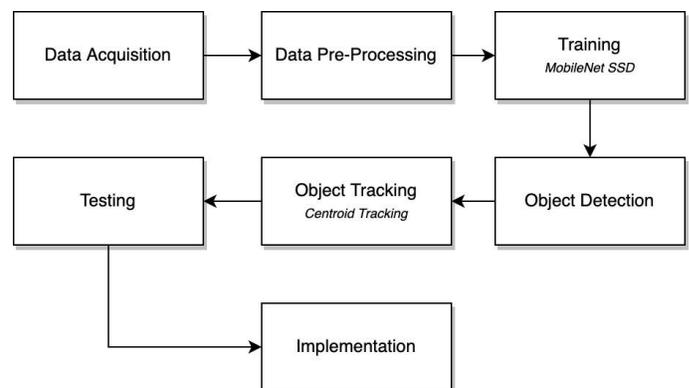


Fig. 1. Research Methodology.

B. Data Pre-processing Stage

After the input dataset was collected, the annotation process was carried out. Annotation is the process of creating labels by providing a bounding box along with the object class name in each image. In this study, the annotations were stored in a file containing information about the object class, each bounding box's coordinates, and its label. Details of the number of datasets are:

- The dataset is an image of a human class object.
- We amassed about 330 data, consisting of 227 training data and 33 test data.
- The number of annotations on the training data is 1,307, while the number of annotations on the test data is 83.

C. Training Stage

Object recognition training is carried out using an annotated image dataset. Then a convolution process is performed using the MobileNet SSD architecture to train the model weights to accurately categorize objects visible on the video camera. In the training stage, iteration was determined by a threshold indication, which takes the form of a loss function value, as seen in Fig. 2. The lower the loss value, the better the developed object detection model.

The SSD MobileNet architecture combines the MobileNet architecture as the base network and the SSD architecture as the detecting network.

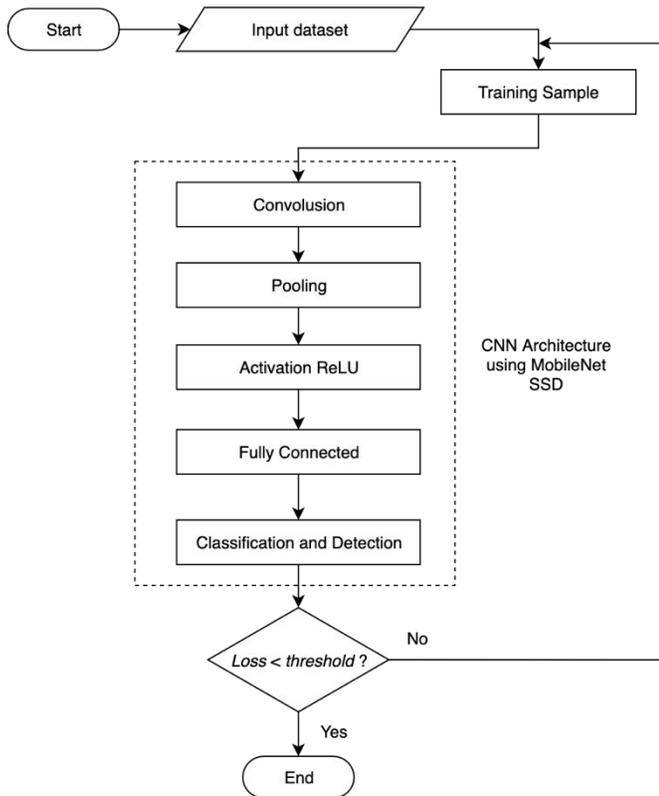


Fig. 2. Block Diagram of Training Process.

Table I depicts MobileNet's standard architecture. Column *n* specifies the number of layers, column *c* specifies the output size, and column *s* specifies the stride [20]. Before average pooling, the MobileNet architecture employs all layers, and the SSD architecture replaces the layers in the typical pooling and fully connected network.

TABLE I. MOBILENET ARCHITECTURE

Input	Operator	t	c	n	s
$224^2 \times 3$	conv2d	-	32	1	2
$112^2 \times 32$	bottleneck	1	16	1	1
$112^2 \times 16$	bottleneck	6	24	2	2
$56^2 \times 24$	bottleneck	6	32	3	2
$28^2 \times 32$	bottleneck	6	64	4	2
$14^2 \times 64$	bottleneck	6	96	3	1
$14^2 \times 96$	bottleneck	6	160	3	2
$7^2 \times 160$	bottleneck	6	320	1	1
$7^2 \times 320$	conv2d 1x1	-	1280	1	1
$7^2 \times 1280$	avgpool 7x7	-	-	1	-
$1 \times 1 \times 1280$	conv2d 1x1	-	k	-	-

A multiscale feature map layer is employed in SSD design, which indicates that various feature map sizes are used. In general, the feature map sizes on SSDs are 512, 256, 256, and 128 [21]. Modifications were performed in this study by adding a 64 feature map, resulting in the SSD network configuration being 512, 256, 256, 128, and 64. The addition of this feature map seeks to evaluate the model's effectiveness in recognizing items of smaller size.

D. Object Detection

After the training stage is completed, the model can be used to identify the trained objects. Fig. 3 depicts the steps of object identification.

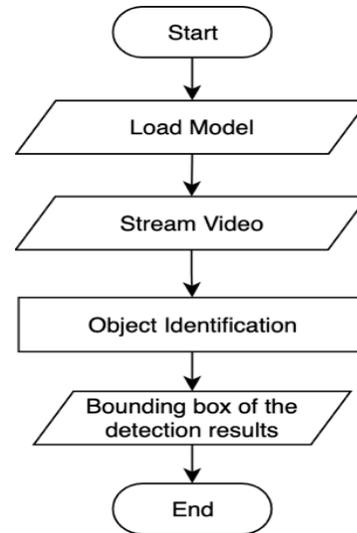


Fig. 3. Block Diagram of Object Identification Process.

E. Object Tracking

Object tracking is used to identify the detecting object so that its movement in a movie can be monitored from frame to frame. The object tracking algorithm used in this work is a centroid tracking approach, although it has been changed to improve performance. Tracking workflow can be seen in Fig. 4. The centroid tracking algorithm is based on the Euclidean distance between the existing centroid object and the new centroid object between frames in a video. The following are the suggested tracking steps:

- Determine the centroid at each of the bounding box locations.
- Calculate the Euclidean distance between the new and old bounding boxes.
- Update the coordinates (x, y) based on the object and class category.
- If a new detection occurs, the new object is added to the new track.
- If the tracking object has been lost, it is removed by setting a threshold for the next N frames.

The centroid tracking approach can perform effectively if the object detection model delivers correct results. The object detection model may not be accurate in the object detection process. Therefore, there is the possibility of mistakes in the form of objects that are not identified or detected but are less accurate. For example, one object is detected as two or more objects. The Non-Maximum Suppression (NMS) method is used in this work to suggest an additional way of post-process detection. NMS works to solve the problem of overlapping bounding boxes from detection results. Incorporating this process will aid the detection process in determining if the bounding box above another bounding box is still the same object or a distinct object [22]. The following are suggested steps from NMS:

- Step 2 should be repeated for each separate object class.
- Take the last index from the index list's enclosing box and add the index value to the specified index list.
- Find the greatest coordinate (x, y) for the bounding box's start and the smallest coordinate (x, y) for the bounding box's end.
- Determine the bounding box's width and height.
- Determine the overlap ratio.
- Remove from the index list any indexes with overlapping threshold values.

F. Testing Stage

The tracking performance testing procedure steps are carried out by applying detection and tracking algorithms on image and video testing with a variety of tests, including camera distance and height testing, frame rates at various video resolutions, and object closures. The F1 score from the confusion matrix is used to calculate performance.

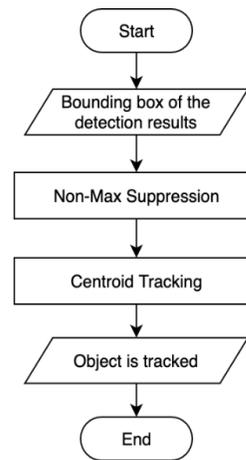


Fig. 4. Block Diagram of Tracking Object Process.

G. Implementation

Following the completion of the testing stages and the development of a good model for detecting human class objects, in this case, the instructor, the implementation is carried out in the classroom.

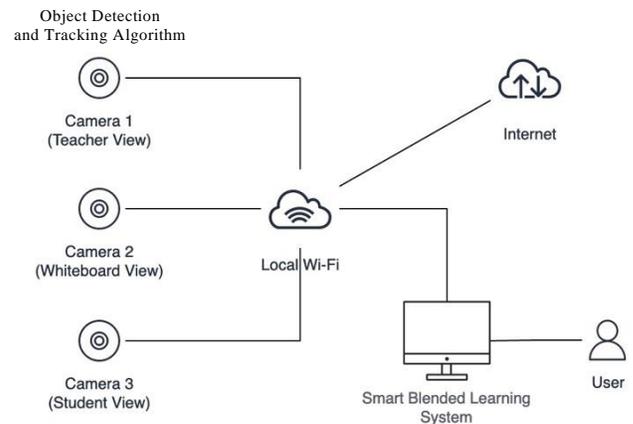


Fig. 5. The Architecture of the Smart Blended Learning System.

The MobileNet SSD object detection algorithm and centroid tracking are only applied to camera 1, pointing to the teacher, as shown in Fig. 5. The algorithm will monitor and zoom in on the teacher, ensuring that the video results in the video conference constantly center on the teacher.

IV. RESULT AND DISCUSSION

The F1 score test is used to evaluate the performance of the object detection algorithm by taking precision and recall into account. In 5,079 steps, the human object class is trained according to the results in Fig. 6. The resulting mAP value is obtained by employing transfer learning techniques, which allow the trained model to converge faster because the initialization of weights in the convolution process does not begin at random but rather uses the weights value of the model that has been trained on a large dataset with a variety of objects. This transfer learning technique enables the model to understand the pattern of each object class being taught more quickly.

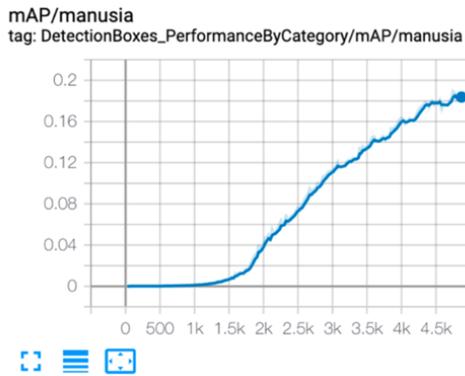


Fig. 6. Mean Average Precision.

The transfer learning technique allows the trained model to converge faster since the initialization of weights in the convolution process does not begin at random but instead employs the weights values of the previously trained model on big datasets with various objects. This transfer learning strategy enables the model to recognize the pattern of the object class being learned more quickly. Fig. 7 depicts the statistics of the loss value produced in each iteration.

Because this study focuses on predicting the positive class rather than the negative class, average precision and recall data are employed; suppose we have a total of 10,000 pieces of data. However, only 40 data points have positive labels, while 9,960 data are classified as negative. By predicting all negative classifications, the algorithm can achieve an accuracy of more than 99 percent with this data composition. The resulting model is less efficient if it can only be accurate in the negative class. Metric precision and recollection are required to address this difficulty. The mAP value varies from 0 to 1, with 0 being the lowest and 1 being the highest. The trained model's final mAP value on box detection performance for the human class is 0.1926.

The loss value is calculated from step to step till the training procedure is terminated. Losses are calculated in three ways: classification loss, localization loss, and regularization loss. Classification loss displays the error value for the object class's classification results. Localization loss is a number that expresses the error in determining the position of an object's bounding box during the inference process. Meanwhile, regularization loss is a process that adjusts or reduces the coefficient to zero, which is vital for assisting the trained network is converging more quickly. Based on Fig. 7, the last classification loss value is 6.02, the last localization loss value is 2.15, the last regularization loss value is 0.346, and if they are added together, the total loss is 8.516. The smaller the loss value, the better the performance of the resulting model.

Furthermore, the model's capacity to detect distance and camera height is tested to see how well the model performs, which will subsequently be utilized to detect things for teachers with varied detection distances and camera installation heights. The distances tested in this test are 2 meters, 4 meters, 6 meters, 8 meters, and 10 meters. While the camera tested has a height of 1.5 meters and 3 meters. Fig. 8 depicts an example of a test with varying distances.

The model testing results with varied distances and camera heights are shown in Table II. The dataset utilized was gathered independently by photographing human things at various distances and heights. The test results utilizing a camera location at the height of 1.5 meters revealed that the model could recognize human objects at detection distances of 2, 4, and 6 meters, as indicated by an F1 score of 1. It was discovered that the model's capacity to identify human objects is good at distances of 2, 4, and 6 meters while at a camera height of 3 meters. Object identification performance decreased significantly at a distance of 10 meters with an F1 score of 0.28.

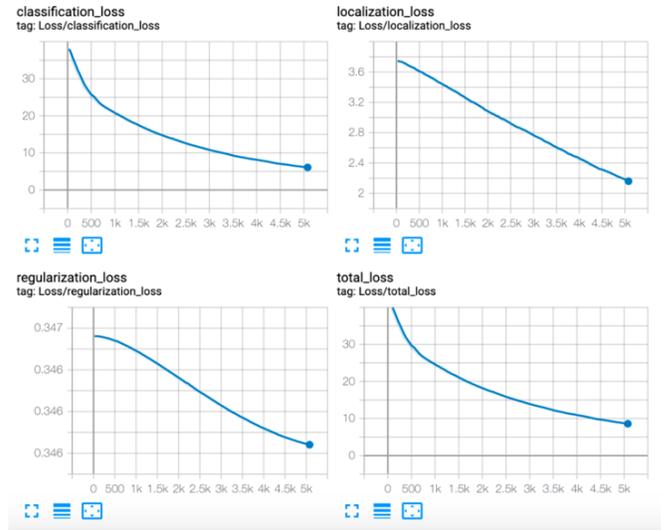


Fig. 7. Statistics and Loss Function Values.

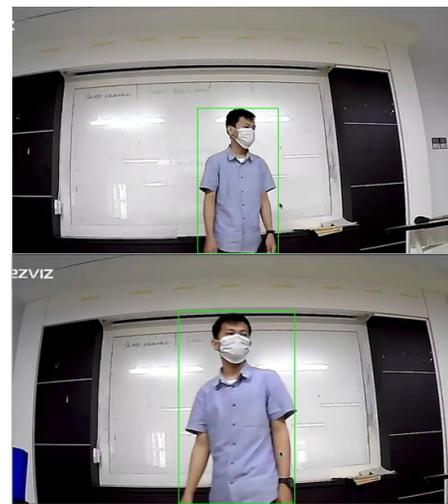


Fig. 8. Detection Results at Different Distances.

TABLE II. F1-MEASURE FOR DIFFERENT DISTANCES AND HEIGHTS OF CAMERA

No	Object Class	Height	Distance				
			2 m	4 m	6 m	8 m	10 m
1	Person	1.5 m	1	1	1	0.75	0.75
		3 m	0.92	1	0.8	0.76	0.28

This study adds new parameters to the detector network layer, specifically a new feature map with a depth of 64 at the network layer's end. The feature map is made smaller than the previous layer, which has a depth of 128. The addition of this layer attempts to provide the trained model the ability to distinguish smaller or farther away objects. The test results following the addition revealed that the detection results were steadier from a distance of 2 meters to 6 meters. However, at a distance of 8 and 10 meters, the accuracy value fell dramatically. Measurement results with varying distances are carried out in a room with sufficient light intensity, and limited lighting sources can affect the results of object detection in the system.

This study success performs object detection and tracking technology to shoot a wide area in the classroom so that wherever the teacher moves, the video will still focus on the teacher in the center of the video. It will provide space for teachers to teach and move freely in the classroom.

V. CONCLUSION AND FUTURE WORK

We created a learning system solution in the form of a smart blended learning system based on a video camera in this framework by adding object detection and tracking. This solution has made online and offline learning more participatory for teachers and students.

So far, we've successfully implemented and piloted a smart blended learning system in a real-world classroom setting. A static camera is utilized in this video. Our next main project will involve the usage of a dynamic camera with a pan and tilt mechanism.

The design and implementation of our smart blended learning system framework are still ongoing, emphasizing the difficulties that continue to emerge and must be addressed to achieve a smart blended learning system framework in a global formulation.

ACKNOWLEDGMENT

The authors are grateful to the Indonesian Ministry of Education, Culture, Research, and Technology for funding this research through the Higher Education Leading Applied Research (Kementerian Pendidikan, Kebudayaan, Riset dan Teknologi or PTUPT) scheme in 2021. We are obliged to Universitas Negeri Makassar (UNM) to provide the facilities and equipment needed for this research to be carried out.

REFERENCES

- [1] A. D. Ekawati, L. Sugandi, and D. L. Kusumastuti, "Blended learning in higher education: Does gender influence the student satisfaction on blended learning?," *Proc. 2017 Int. Conf. Inf. Manag. Technol. ICIMTech 2017*, vol. 2018-Janua, no. November, pp. 160–164, 2018, DOI: 10.1109/ICIMTech.2017.8273530.
- [2] A. Al-Hunaiyyan and S. Al-Sharhan, "The design of multimedia blended e-learning systems: Cultural considerations," *3rd Int. Conf. Signals, Circuits Syst. SCS 2009*, pp. 1–5, 2009, DOI: 10.1109/ICSCS.2009.5412342.
- [3] B. Zhang, "Computer vision vs. human vision," in *9th IEEE International Conference on Cognitive Informatics (ICCI'10)*, 2010, p. 3. DOI: 10.1109/COGINF.2010.5599750.

- [4] A. Wahid, S. Luhriyani, Nurhikmah, J. M. Parenreng, M. F. B, and M. I. Nur, "Smart Campus Framework: A Solution for New Normal Education System," in *2021 IEEE 5th International Conference on Information Technology, Information Systems and Electrical Engineering (ICITISEE)*, 2021, pp. 266–271. DOI: 10.1109/ICITISEE53823.2021.9655952.
- [5] J. Yang, H. Yu, and N.-S. Chen, "Using blended synchronous classroom approach to promote learning performance in rural area," *Comput. Educ.*, vol. 141, p. 103619, Jul. 2019, DOI: 10.1016/j.compedu.2019.103619.
- [6] M. Hastiea, I. C. Hung, N. S. Chen, and Kinshuk, "A blended synchronous learning model for educational international collaboration," <https://doi.org/10.1080/14703290903525812>, vol. 47, no. 1, pp. 9–24, Feb. 2010, DOI: 10.1080/14703290903525812.
- [7] J. Carman, "Blended learning design: Five key ingredients," p. 11, Jan. 2005.
- [8] M. Abisado, "A Flexible Learning Framework Implementing Asynchronous Course Delivery for Philippine Local Colleges and Universities," *Int. J. Adv. Trends Comput. Sci. Eng.*, vol. 9, pp. 413–421, Jun. 2020, DOI: 10.30534/ijatcse/2020/6591.32020.
- [9] Q. Wang, C. L. Quek, and X. Hu, "Designing and Improving a Blended Synchronous Learning Environment: An Educational Design Research," *Int. Rev. Res. Open Distrib. Learn.*, vol. 18, no. 3, pp. 99–118, May 2017, DOI: 10.19173/IRRODL.V18I3.3034.
- [10] A. Yilmaz, O. Javed, and M. Shah, "Object tracking: A survey," *ACM Comput. Surv.*, vol. 38, no. 4, 2006, DOI: 10.1145/1177352.1177355.
- [11] B. Zhong et al., "Visual tracking via weakly supervised learning from multiple imperfect oracles," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, no. July 2010, pp. 1323–1330, 2010, DOI: 10.1109/CVPR.2010.5539816.
- [12] A. Salhi and A. Y. Jammoussi, "Object tracking system using Camshift, MeanShift and Kalman filter," *World Acad. Sci. Eng. Technol.*, vol. 6, no. 4, pp. 674–679, 2012.
- [13] Z. Rahman, A. M. Ami, and M. A. Ullah, "A Real-Time Wrong-Way Vehicle Detection Based on YOLO and Centroid Tracking," *2020 IEEE Reg. 10 Symp. TENSYPMP 2020*, no. June, pp. 916–920, 2020, DOI: 10.1109/TENSYPMP50017.2020.9230463.
- [14] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, pp. 580–587, 2014, DOI: 10.1109/CVPR.2014.81.
- [15] R. Girshick, "Fast R-CNN," *Proc. IEEE Int. Conf. Comput. Vis.*, vol. 2015 Inter, pp. 1440–1448, 2015, DOI: 10.1109/ICCV.2015.169.
- [16] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, 2017, DOI: 10.1109/TPAMI.2016.2577031.
- [17] Y. Li, H. Huang, Q. Xie, L. Yao, and Q. Chen, "Research on a Surface Defect Detection Algorithm Based on MobileNet-SSD," *Appl. Sci.*, vol. 8, no. 9, p. 1678, Mar. 2018, DOI: 10.3390/app8091678.
- [18] A. G. Howard et al., "MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications," *arXiv1704.04861 [cs]*, Feb. 2017, [Online]. Available: <http://arxiv.org/abs/1704.04861>.
- [19] A. Alimudin and A. F. Muhammad, "Online video conference system using WebRTC technology for distance learning support," *Int. Electron. Symp. Knowl. Creat. Intell. Comput. IES-KCIC 2018 - Proc.*, pp. 384–387, 2019, doi: 10.1109/KCIC.2018.8628568.
- [20] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L. C. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," *arXiv*, 2018.
- [21] W. Liu et al., "SSD: Single Shot MultiBox Detector," *arXiv1512.02325 [cs]*, vol. 9905, pp. 21–37, Jan. 2016, doi: 10.1007/978-3-319-46448-0_2.
- [22] N. Bodla, B. Singh, R. Chellappa, and L. S. Davis, "Soft-NMS -- Improving Object Detection With One Line of Code," *arXiv1704.04503 [cs]*, Jun. 2017, [Online]. Available: <http://arxiv.org/abs/1704.04503>.

Smart Agriculture Monitoring System using Clean Energy

Karim ABOUELMEHDI, Kamal ELHATTAB, Abdelmajid EL MOUTAOUAKKIL
LAROSERI Laboratory, FS Chouaib Doukkali University
El Jadida, Morocco

Abstract—Internet of Things (IoT) technology makes all areas of human life more comfortable. The development of farms through the use of IoT positively influences agricultural production not only by strengthening it, but also by making it more profitable and reducing the cost of production. The goal of this paper is to offer a new IoT-based smart agriculture system that helps farmers get real-time data such as (temperature, humidity, soil moisture) for effective environmental monitoring that will allow them to increase overall yield and product quality. The farm monitoring system proposed in this paper is based on the ESP32 microcontroller with a set of sensors. This new model produces a real-time data feed that can be viewed online via a mobile app. The proposed new system uses solar energy with battery as an energy source.

Keywords—IoT; smart agriculture; new model; solar panel; esp32; mobile app

I. INTRODUCTION

Agriculture has always played a very important role for the human being and for the economies of countries, indeed countries that have experienced significant economic growth have experienced a significant increase in the agricultural sector, agriculture is not only an important factor of economic growth but also a heritage, a way of life, a cultural identity of a country. The use of the Internet of Things in agriculture makes it possible to put technology at the service of the development of agriculture and also to help the farmer to:

- Make very important decisions at the right times.
- Save energy and time.
- Deal with the lack of information.
- Increase agricultural production.
- Allow the farmer to develop and develop his agricultural production.
- Increase the participation of agriculture in job creation.
- Ensure our country's needs for agricultural products.

The organization of the article is as follows, in Section II (Literature Survey): we collected the work carried out in the field of smart agriculture, Section III (discussions and results): we treated and analyzed the work done and proposed improvements in this area, Section IV (the new model): we proposed our new intelligent farm system, and Section V (Conclusion and outlook).

II. LITERATURE SURVEY

To study the use of IOT in the agricultural field, it is essential to make a literary study on the current state, projects and studies that have been carried out in this field, this step is important to be able to have visibility on the current state and on the difficulties encountered, it will allow us to make proposals for improvement in this area.

A. Data Collection

To carry out our literary study we focused our research work on three databases ACM Library, Science Direct and IEEE. The Pattern used is IOT and agriculture during the last four years (2018-2022) this has allowed us to have 29 publications on the different methods and procedures of use of the IOT in the agricultural field as well as the advantages, disadvantages, and prospects for improvement in this field, Table I below summarizes the research approach.

TABLE I. ARTICLES

Database	Search pattern	Years of publication	Number of articles
ACM Library	IOT and agriculture	(2018-2022)	8
Science Direct	IOT and agriculture	(2018-2022)	10
IEEE	IOT and agriculture	(2018-2022)	12

TABLE II. DATA COLLECTION (SMART AGRICULTURE)

Year / Author	Sous-domain	Challenges	Data / Sensors	Technologies
Hsiao-Tzu (2018) [1]	✓ Agriculture monitoring	✓ Monitor the condition of agricultural land	✓ Temperature ✓ Humidity ✓ CO2	✓ Wi-Fi ✓ Node MCU ✓ Rasepberr y ✓ Web technology
Joy G. Bea (2019) [2]	✓ Smart farming	✓ Monitor a chicken farm	✓ Temperature ✓ Humidity ✓ Ammonia ✓ Ultra sonic	✓ Arduino ✓ Wifi ✓ Coud technology ✓ Mobile technology
Waleed	✓ Greenhou	✓ Examine	✓ Soil	✓ Arduino

(2018) [23]	Monitoring	parameters using IOT.	✓Humidity ✓Temperature ✓Light	✓Cloud ✓Mobile ✓Wifi
Teddy surya (2019) [23]	✓Smart farming	✓Develop an smart chicken poultry farm.	✓Humidity ✓Temperature ✓Ammonia ✓CO2	✓Arduino ✓Wifi
Jishakc (2018) [24]	✓Agriculture Monitoring	✓IOT-Based water level monitoring.	✓Moisture ULTRASONIC	✓Arduino ✓Cloud ✓Mobile
M Manideep (2019) [25]	✓Agriculture Monitoring	✓Smart agriculture with image capture module.	✓Soil Moisture ✓Humidity ✓Temperature	✓Arduino ✓Camera ✓GSM
Jenny priyanka (2020) [26]	✓Agriculture Monitoring	✓Develop an smart monitoring system for poultry farm.	✓Humidity ✓Temperature	✓Nodmcu ✓Cloud ✓Mobile ✓SMS
R Nageswara rao (2018) [27]	✓Smart irrigation	✓Monitor crops with a automated irrigation system	✓Soil Moisture ✓Temperature	✓Raspeberry ✓Cloud technology ✓Mobile technology
Bilgi gorkem yazgac (2021) [28]	✓Agriculture Monitoring	✓Develop a monitoring system for agriculture.	✓Soil Moisture ✓Temperature	✓Nodmcu ✓Cloud ✓Web technology
Fan zhany (2020) [29]	✓Agriculture Monitoring	✓Develop an intelligent green house management system.	✓Humidity ✓Temperature ✓Winddirection ✓Windspeed ✓Light	✓Nodmcu ✓Cloud ✓Web technology ✓Bluetooth

B. Data Processing

The processing and analysis of the work carried out in this area will allow us to compare the following attributes: data collected, technologies used, challenges of the current approach. Table II shows this comparison. During our analysis, we excluded the number of sensors, the amount of data collected, the underlying technologies, the topology of the sensors and other intermediate gateways were not included as this does not add value to our research work.

III. DISCUSSIONS AND RESULTS

A. Data Analysis

The analysis of the data gave us the opportunity to have quantitative and qualitative data on the use of IOT in the agricultural field, our analysis focuses on the agricultural field and the types of data collected, the technologies used and the areas of application. The objective of this analysis is to increase productivity and efficiency while saving natural and human resources in the agricultural sector.

B. Agricultural Domains

We have grouped the articles and classified them by agricultural field, as shown in Table III below. This ranking gave us four agricultural sub-domains, following the results obtained we deduced that smart irrigation is the most treated agricultural field with a percentage of 40%.

TABLE III. AGRICULTURAL SUB-DOMAINS

Agricole domain	Percentage (%)
Smart irrigation	(40%)
Agriculture monitoring	(33%)
Smart farming	(17%)
Greenhouse agriculture	(10%)

C. Data Collected

Data collection is part of the objectives of the IOT, so it is necessary to collect a large amount of data to give accurate results, to follow up on our analysis it was deduced that many articles are focused on temperature (31%) and soil moisture (28%), and humidity (25%). Other data were also collected but with a small percentage. Table IV shows the results.

TABLE IV. DATA (SENSORS)

Data collected	Percentage (%)
Temperature	(31%)
Soil Moisture	(28%)
Humidity	(25%)
Light	(5%)
Ph	(4%)
Water level	(4%)
CO2	(3%)

D. Technologies used

The grouping of technologies used in the articles covered tells us that cloud technology is identified as the most used technology (23%), followed by Arduino and Mobile technologies (16%). Followed by WIFI (15%), followed by Nodmcu (12%) and followed by Raspeberry (11%) other technologies were also used but with a small percentage. Table V shows the results.

TABLE V. TECHNOLOGY

Technology	Percentage (%)
Cloud technology (23%)	(23%)
Wifi (15%)	(15%)
Arduino (16%)	(16%)
Mobile technology (16%)	(16%)
Nodmcu (12%)	(12%)
Raspeberry (11%)	(11%)
Web technology (2%)	(2%)

E. Results

In this paper, we were able to study the use of IOT in the agricultural field, this allowed us to identify important attributes to analyze it, for this we gathered and analyzed recent scientific data, this survey was a way to have the list of the most studied sub-areas: Smart irrigation, Agriculture Monitoring, Smart farming, Greenhouse Agriculture.

Smart irrigation is the most studied agricultural sub-field in recent years, as most countries focus mainly on the use of water resources due to its lack. Smart irrigation has become an essential means that positively influence the agricultural field and its production, current efforts focus on the management of water resources this has given great value to irrigation management to increase the quality and quantity of agricultural products.

The second sub-area is forecast agriculture because the demand for products is constantly increasing so it is essential to opt for a forward-looking management of agricultural production. This makes it possible to review any risk that can block the objectives of farmers and makes it possible to ensure a quality and increase in agricultural production and also saves resources. Based on our analysis, it was found that ambient temperature followed by soil moisture and moisture are among the most commonly measured data. According to the results of our study, it was found that the use of the Internet of Things in agriculture can be used effectively to increase agricultural production to meet the growing needs of the population.

The majority of research focuses on water management by monitoring environmental parameters such as temperature, humidity and soil moisture. Many results have emphasized the importance of the proper management of water resources, saving human and material resources.

Innovation in the IT field makes it possible to cover all agricultural areas to allow an important development. IoT has solved many problems related to agriculture and farming, but several limitations are to be taken into consideration such as the lack of interoperability and compatibility of devices, the problems of network flexibility when several devices connect, and the lifetime of the sensor are some of the limitations to be solved in future research.

Food production is a great challenge for all economies around the world due to the rapid and constant evolution of the world's population, especially with climate change and labour shortages. Currently, current research focuses on robotics to solve these problems. Many researchers and companies have focused on robotics and artificial intelligence (AI) to reduce the amount of herbicide used by farmers.

Following the analysis of the data collected, it was found that the IOT is in great demand to develop the agricultural field, several researches have been carried out in relation to this subject, the articles that we studied in this report have not taken into consideration the sources of energy to power the IOT equipment and the security of agricultural fields against intrusions, thus the majority of articles use Arduino technology for the interconnection of IOT equipment, which is why we will propose a new model that will take into account the energy

source, uses an interconnection technology with low energy consumption and secures agricultural fields against intrusions.

IV. NEW MODEL

A. Solution

After the in-depth study of the work done in this area, we found that agriculture is one of the main areas that use IoT. Several IoT components have already been put in place to serve the agricultural field to monitor soil conditions such as temperature, moisture, soil moisture and soil ph.

We also found that most authors propose solutions provided by wires or batteries such as power sources, the things that make powering objects a real challenge for several reasons, the most important of which are:

- The area of agricultural land is large, which increases the cost of installing electrical cables. The installed cables remain subject to damage that can be caused by agricultural vehicles, insects, and water, as the cables are underground.
- The difficulties of maintenance and repair in case of failure because the cables are underground it is necessary at each failure to remove them to repair them.
- Battery life remains limited and varies from a few hours to several years.
- The power of objects by alkaline batteries has major disadvantages. These batteries have a limited amount of energy.
- The cost of purchasing and replacing the batteries of the objects.

We offer a solution that offers an improvement of existing solutions. Our model is highly customizable and provides a data analysis solution that enables large-scale data processing on real-time observation streams. Our new model uses solar energy with battery as a power source, the ESP32 microcontroller as a low-energy interconnection technology [30] and a motion detector to protect our agricultural field from intrusions.

B. Components

1) *Ph: Sensor:* The PH sensor is a sensor that allows determining the acidity of the soil between 0 and 14 at a temperature between 0 and 60 ° C, this acidity varies depending on the availability of nutrients in water / soil. The detection of PH is therefore essential to understand how to fertilize it properly in order to prepare a good environment for agricultural production. Fig. 1 shows the ph sensor.

2) *Temperature and humidity sensor:* The DHT11 sensor is a low-cost basic digital temperature and humidity sensor. Capable of measuring temperatures from 0 to 50°C with an accuracy of 2°C and relative humidity levels from 20 to 80% with an accuracy of 5%. A measurement can be made every second. Fig. 2 shows the DHT11 sensor.



Fig. 1. Ph Sensor.



Fig. 2. Temperature and Humidity Sensor.

3) *PIR Sensor*: The passive infrared sensor is used for intrusion detection in agricultural fields and helps farmers detect movements (of animals and humans) in their agricultural fields. Fig. 3 shows the PIR sensor.



Fig. 3. PIR Sensor.

4) *Soil Moisture Sensor*: This sensor, as shown in Fig. 4, measures soil moisture due to changes in the electrical conductivity of the earth (soil resistance increases with drought). A digital output with an adjustable threshold is used to trigger an alarm or sprinkler pump. A second analog output accurately tracks fluctuations in soil moisture. When the sensor fork is planted vertically in the ground (agricultural fields, garden, etc.). The electrical resistance between the two electrodes is measured.



Fig. 4. Soil Moisture Sensor.

5) *ESP32 Microcontroller*: The ESP32, illustrated in Fig. 5 designed and produced by the company Espressif, it integrates functions dedicated to the Internet of Things and more particularly wireless communications and Bluetooth and lora for a reduced cost. It is the best choice for smart cities, smart farms, and smart homes.



Fig. 5. Esp32.

6) *Solar panel*: Solar panels, shown in Fig. 6, are the devices used to absorb the sun's rays and convert them into electricity. In our case we will use solar panels to power our IoT devices.

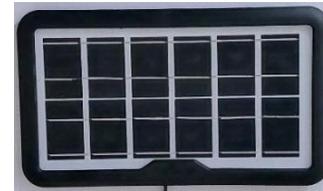


Fig. 6. Solar Panel.

7) *Lithium Batteries*: In our case we will use the lithium battery illustrated in Fig. 7, to have the possibility of setting up an autonomous network equipped with a battery capable of storing energy during the day and provides electricity during the night.



Fig. 7. Batteries.

8) *Top-up card for power bank*: In our case we will use the charge card, illustrated in Fig. 8, to build a power bank from lithium battery cells to charge all IOT devices compatible with 5V.



Fig. 8. Power Bank.

C. System Description

To achieve a high-quality yield, we will build a new smart farm monitoring system powered by solar energy. The proposed solution consists of two parts, a hardware part and a logicielle. The hardware part consists of a transmitter node that contains the pH, soil moisture, temperature, humidity, and motion sensor connected to ESP32 and a solar panel and battery to supply our system with electrical energy. The software part consists of a mobile application that allows farmers to monitor their agricultural environment as long as it

is connected to the internet. The method of data acquisition is illustrated in Fig. 9. The pH, soil moisture, temperature, humidity, and motion sensors send digital data to the ESP32. This data will be transmitted to the cloud server via WIFI communication. When there is incoming data on the cloud server, farmers can monitor their farming environment via a mobile app.

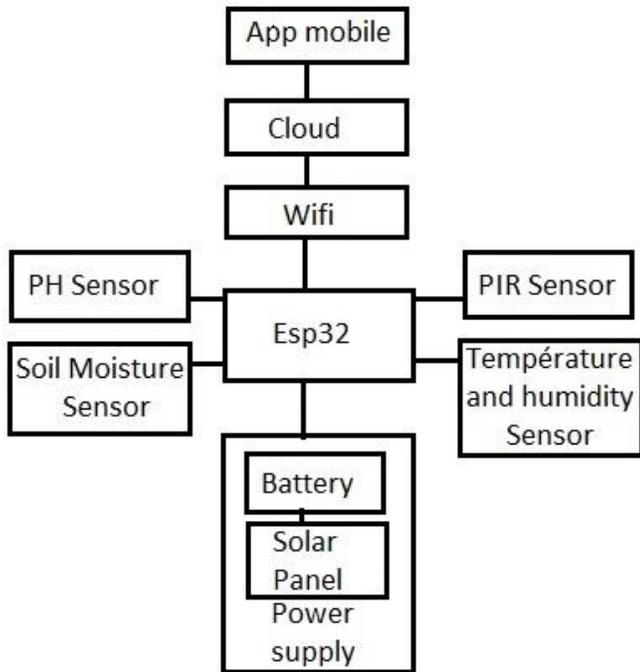


Fig. 9. New System.

D. Tools and Libraries used

In this paper we used several software and libraries to implement our new farm monitoring system:

1) *Arduino*: The Arduino integrated development environment is a cross-platform application for Windows, MacOS, and Linux that is written in C and C++ functions. It is used to write and download programs to Arduino-compatible boards. The EDI source code is released under the GNU General Public License, version 2. The IDE environment contains mainly two basic parts: the editor and the compiler where the old one is used to write the required code and later is used to compile and download the code in the given Arduino module.

2) *Libraries*: The Arduino IDE can be extended using libraries, just like most programming platforms. Libraries offer additional features, such as working with hardware or manipulating data. The standard libraries used are described below.

a) *DHT.h*: To read the data from the DHT sensor, we will use Adafruit's DHT.h library. To use this library, we must also install the Adafruit Unified Sensor library.

b) *WiFi.h*: This library allows our esp32 card to connect to the Wi-Fi network to start sending and receiving data.

3) *Algorithms and flowchart*: This section presents the algorithms and flowchart of the overall process of our new Smart Farm System:

a) *Algorithms*: Generic process algorithm for temperature, humidity, soil moisture and pH sensors.

STEP 1: start the process;

STEP 2: connected to wifi;

STEP 3: read temperature, humidity, soil and ph;

STEP 4: get temperature, humidity, soil and ph values from analog pins;

STEP 5: send data to cloud;

STEP 6: delay to 10 seconds;

STEP 7: repeat step 4, 5 & 6 until the process end;

STEP 8: end;

b) *Flowchart*: Generic process flowchart for pir sensor motion sensor (Fig. 10).

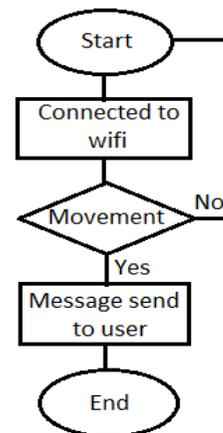


Fig. 10. Flowchart (Pir Sensor).

E. Mobile Application

In creating our app, we used the blynk platform that supports Android. It works with several types of microcontrollers such as Arduino, Raspberry and ESP32. It consists of three main components:

- The blynk app that allows you to control a device and view data.
- The blynk server, which is a cloud service responsible for all communications between objects and the smartphone.
- Blynk libraries, which include various widgets such as control buttons, display formats, and notifications.

When creating our application, as shown in Fig. 11, we tried to make it very simple, with a single interface that shows the main information collected by our ESP32. This application immediately requests real-time information from the cloud to display it.



Fig. 11. Mobile App.

F. System Test

We tested the consumption of the solar power of our new model for seven days. Using the display screen of our new model shown in Fig. 12, we noticed that with the equipment used for four days, the consumption of our battery does not exceed 26%.

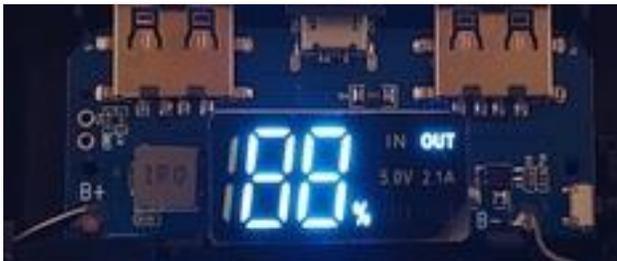


Fig. 12. Test.

V. CONCLUSION AND PERSPECTIVES

In this article, a new smart farm monitoring model was realized. To propose this model, we started by studying the work already done in the field of smart agriculture, the analysis of the data allowed us to raise three major questions: The energy sources to power the equipment, the technologies used to interconnect these objects and the security of agricultural fields against intrusions. The tests we carried out for our new agricultural field monitoring system were positive, using new devices and sensors for example (PIR sensor).

The prospects of our project are to test our new model in a large agricultural field to study the autonomy of the batteries connected to the solar panels and propose improvements to our new system.

REFERENCES

[1] H.-T. Hsu, T.-M. Wang, et Y.-C. Kuo, « Implementation of Agricultural Monitoring System Based On The Internet of Things », in Proceedings of the 2018 2nd International Conference on Education and E-Learning, Bali Indonesia, nov. 2018, p. 212-216. doi: 10.1145/3291078.3291098.

[2] J. G. Bea et J. S. D. Cruz, « Chicken Farm Monitoring System Using Sensors and Arduino Microcontroller », in Proceedings of the 9th International Conference on Information Systems and Technologies, Cairo Egypt, mars 2019, p. 1-4. doi: 10.1145/3361570.3361607.

[3] W. Abdallah, M. Khair, M. Ayyash, et I. Asad, « IoT system to control greenhouse agriculture based on the needs of Palestinian farmers », in Proceedings of the 2nd International Conference on Future Networks and Distributed Systems, Amman Jordan, juin 2018, p. 1-9. doi: 10.1145/3231053.3231061.

[4] K.-M. Chew, S. C.-W. Tan, G. C.-W. Loh, N. Bundan, et S.-P. Yitong, « IoT Soil Moisture Monitoring and Irrigation System Development », in Proceedings of the 2020 9th International Conference on Software and Computer Applications, Langkawi Malaysia, févr. 2020, p. 247-252. doi: 10.1145/3384544.3384595.

[5] D. Mishra, T. Pande, K. K. Agrawal, A. Abbas, A. K. Pandey, et R. S. Yadav, « Smart agriculture system using IoT », in Proceedings of the Third International Conference on Advanced Informatics for Computing Research - ICAICR '19, Shimla, India, 2019, p. 1-7. doi: 10.1145/3339311.3339350.

[6] P. Netinam, A. Niratsoke, et M. Rukhiran, « Beyond Traditional Piggery to Automation Farm System Based on Internet of Things », in 2021 The 5th International Conference on E-Commerce, E-Business and E-Government, Rome Italy, avr. 2021, p. 39-42. doi: 10.1145/3466029.3466040.

[7] J. Sun, A. M. Abdulghani, M. A. Imran, et Q. H. Abbasi, « IoT Enabled Smart Fertilization and Irrigation Aid for Agricultural Purposes », in Proceedings of the 2020 International Conference on Computing, Networks and Internet of Things, Sanya China, avr. 2020, p. 71-75. doi: 10.1145/3398329.3398339.

[8] R. Zhao, S. Ma, et Y. Ding, « Greenhouse Monitoring System Based on Android Platform », in Proceedings of the 2018 2nd International Conference on Big Data and Internet of Things - BDIOT 2018, Beijing, China, 2018, p. 153-156. doi: 10.1145/3289430.3289444.

[9] A. Pathak, M. AmazUddin, Md. J. Abedin, K. Andersson, R. Mustafa, et M. S. Hossain, « IoT based Smart System to Support Agricultural Parameters: A Case Study », Procedia Computer Science, vol. 155, p. 648-653, 2019, doi: 10.1016/j.procs.2019.08.092.

[10] J. Doshi, T. Patel, et S. kumar Bharti, « Smart Farming using IoT, a solution for optimally monitoring farming conditions », Procedia Computer Science, vol. 160, p. 746-751, 2019, doi: 10.1016/j.procs.2019.11.016.

[11] E. A. Abioye et al., « IoT-based monitoring and data-driven modelling of drip irrigation system for mustard leaf cultivation experiment », Information Processing in Agriculture, vol. 8, no 2, p. 270-283, juin 2021, doi: 10.1016/j.inpa.2020.05.004.

[12] A. Subeesh et C. R. Mehta, « Automation and digitization of agriculture using artificial intelligence and internet of things », Artificial Intelligence in Agriculture, vol. 5, p. 278-291, 2021, doi: 10.1016/j.aiaa.2021.11.004.

[13] N. K. Nawandar et V. R. Satpute, « IoT based low cost and intelligent module for smart irrigation system », Computers and Electronics in Agriculture, vol. 162, p. 979-990, juill. 2019, doi: 10.1016/j.compag.2019.05.027.

[14] A. Vij, S. Vijendra, A. Jain, S. Bajaj, A. Bassi, et A. Sharma, « IoT and Machine Learning Approaches for Automation of Farm Irrigation System », Procedia Computer Science, vol. 167, p. 1250-1257, 2020, doi: 10.1016/j.procs.2020.03.440.

[15] H. Benyezza, M. Bouhedda, et S. Rebouh, « Zoning irrigation smart system based on fuzzy control technology and IoT for water and energy saving », Journal of Cleaner Production, vol. 302, p. 127001, juin 2021, doi: 10.1016/j.jclepro.2021.127001.

[16] A. Goap, D. Sharma, A. K. Shukla, et C. Rama Krishna, « An IoT based smart irrigation management system using Machine learning and open source technologies », Computers and Electronics in Agriculture, vol. 155, p. 41-49, déc. 2018, doi: 10.1016/j.compag.2018.09.040.

[17] A. K. Podder et al., « IoT based smart agrotech system for verification of Urban farming parameters », Microprocessors and Microsystems, vol. 82, p. 104025, avr. 2021, doi: 10.1016/j.micpro.2021.104025.

- [18] M. M. Anghelof, G. Suci, R. Craciunescu, et C. Marghescu, « Intelligent System for Precision Agriculture », in 2020 13th International Conference on Communications (COMM), Bucharest, Romania, juin 2020, p. 407-410. doi: 10.1109/COMM48946.2020.9141981.
- [19] M. M. Abbassy et W. M. Ead, « Intelligent Greenhouse Management System », in 2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS), Coimbatore, India, mars 2020, p. 1317-1321. doi: 10.1109/ICACCS48705.2020.9074345.
- [20] R. K. Jain, B. Gupta, M. Ansari, et P. P. Ray, « IOT Enabled Smart Drip Irrigation System Using Web/Android Applications », in 2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT), Kharagpur, India, juill. 2020, p. 1-6. doi: 10.1109/ICCCNT49239.2020.9225345.
- [21] L. Raju K. et V. Vijayaraghavan, « IoT and Cloud hinged Smart Irrigation System for Urban and Rural Farmers employing MQTT Protocol », in 2020 5th International Conference on Devices, Circuits and Systems (ICDCS), Coimbatore, India, mars 2020, p. 71-75. doi: 10.1109/ICDCS48716.2020.243551.
- [22] R. K. Kodali, S. Yerroju, et S. Sahu, « Smart Farm Monitoring Using LoRa Enabled IoT », in 2018 Second International Conference on Green Computing and Internet of Things (ICGCIoT), Bangalore, India, août 2018, p. 391-394. doi: 10.1109/ICGCIoT.2018.8753086.
- [23] V. D. Bachuwar, U. R. Ghodake, A. Lakhssassi, et S. S. Suryavanshi, « WSN/Wi-Fi Microchip-Based Agriculture Parameter Monitoring using IoT », in 2018 International Conference on Smart Systems and Inventive Technology (ICSSIT), Tirunelveli, India, déc. 2018, p. 214-219. doi: 10.1109/ICSSIT.2018.8748638.
- [24] J. Rc, « IoT based Water Level Monitoring and Implementation on Both Agriculture and Domestic.
- [25] M. Manideep, R. Thukaram, et S. M., « Smart Agriculture Farming with Image Capturing Module », in 2019 Global Conference for Advancement in Technology (GCAT), BANGALURU, India, oct. 2019, p. 1-5. doi: 10.1109/GCAT47503.2019.8978368.
- [26] J. P. Mondol, K. R. Mahmud, M. G. Kibria, et A. K. Al Azad, « IoT based Smart Weather Monitoring System for Poultry Farm », in 2020 2nd International Conference on Advanced Information and Communication Technology (ICAICT), Dhaka, Bangladesh, nov. 2020, p. 229-234. doi: 10.1109/ICAICT51780.2020.9333535.
- [27] R. N. Rao et B. Sridhar, « IoT based smart crop-field monitoring and automation irrigation system », in 2018 2nd International Conference on Inventive Systems and Control (ICISC), Coimbatore, janv. 2018, p. 478-483. doi: 10.1109/ICISC.2018.8399118.
- [28] B. G. Yazgac, H. Durmus, M. Kirci, E. O. Gunes, et H. B. Karli, « Petri nets based procedure of hardware/software codesign of an urban agriculture monitoring system », in 2019 8th International Conference on Agro-Geoinformatics (Agro-Geoinformatics), Istanbul, Turkey, juill. 2019, p. 1-4. doi: 10.1109/Agro-Geoinformatics.2019.8820255.
- [29] F. Zhang, X. Wan, T. Zheng, J. Cui, X. Li, et Y. Yang, « Smart Greenhouse Management System based on NB-IoT and Smartphone », in 2020 17th International Joint Conference on Computer Science and Software Engineering (JCSSE), Bangkok, Thailand, nov. 2020, p. 36-41. doi: 10.1109/JCSSE49651.2020.9268351.
- [30] A. Zare et M. T. Iqbal, « Low-Cost ESP32, Raspberry Pi, Node-Red, and MQTT Protocol Based SCADA System », in 2020 IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS), Vancouver, BC, Canada, sept. 2020, p. 1-5. doi: 10.1109/IEMTRONICS51293.2020.9216412.

RENTAKA: A Novel Machine Learning Framework for Crypto-Ransomware Pre-encryption Detection

Wira Z. A. Zakaria¹, Mohd Faizal Abdollah², Othman Mohd³
S. M. Warusia Mohamed S. M. M Yassin⁴, Aswami Ariffin⁵

MyCERT, Cybersecurity Malaysia, Menara Cyber Axis, Jalan Impact, 63000 Cyberjaya, Selangor, Malaysia¹
Faculty of Information and Communication Technology, Universiti Teknikal Malaysia Melaka, Malaysia^{2,3,4}

Hang Tuah Jaya, 76100 Durian Tunggal, Melaka, Malaysia^{2,3,4}

Cyber Security Response Services, Cybersecurity Malaysia, MyCERT, Cybersecurity Malaysia⁵
Menara Cyber Axis, Jalan Impact, 63000 Cyberjaya, Selangor, Malaysia⁵

Abstract—Crypto ransomware is malware that locks its victim's file for ransom using an encryption algorithm. Its popularity has risen at an alarming rate among the cyber community due to several successful worldwide attacks. The encryption employed had caused irreversible damage to the victim's digital files, even when the victim chose to pay the ransom. As a result, cybercriminals have found ransomware a lucrative and profitable cyber-extortion approach. The increasing computing power, memory, cryptography, and digital currency advancement have caused ransomware attacks. It spreads through phishing emails, encrypting sensitive data, and causing harm to the designated client. Most research in ransomware detection focuses on detecting during the encryption and post-attack phase. However, the damage done by crypto-ransomware is almost impossible to reverse, and there is a need for an early detection mechanism. For early detection of crypto-ransomware, behavior-based detection techniques are the most effective. This work describes RENTAKA, a framework based on machine learning for the early detection of crypto-ransomware. The features extracted are based on the phases of the ransomware lifecycle. This experiment included five widely used machine learning classifiers: Naïve Bayes, kNN, Support Vector Machines, Random Forest, and J48. This study proposed a pre-encryption detection framework for crypto-ransomware using a machine learning approach. Based on our experiments, support vector machines (SVM) performed with the best accuracy and TPR, 97.05% and 0.995, respectively.

Keywords—Ransomware; crypto-ransomware; ransomware early detection; pre-encryption; pre-attack; ransomware lifecycle

I. INTRODUCTION

Ransomware is a relatively new type of malware that targets users in an attempt to extort money. Ransomware is malware that encrypts or locks files on an infected computer and demands payment to unlock and decrypt the files. Ransomware was a relatively new intrusion attack that used encryption to extort money from its victim. The victim must follow the ransom note's instructions to pay the ransom in Bitcoin to decrypt and recover the original files. The attacker frequently uses Bitcoin due to its anonymity, as its identity is difficult to trace. On the other hand, paying the ransom does not guarantee that the victim will receive the decryption key necessary to recover the files [1], [2].

Ransomware employs a variety of attack vectors, including social engineering, spam email, botnets, detection evasion, and self-propagation via vulnerabilities. After successfully infecting the victim's machines, it will lock files and directories and encrypt files with the following extensions: .docx, .xlsx, .odt, .zip, .pdf and .jpg. As a result, the victim cannot access their files or computer until the attacker receives the ransom payment within a specified time [3], [4].

Ransomware attacks have grown in sophistication, posing a significant threat to education, health, business, and government organizations. Cybercriminals created hundreds of ransomware variants as a result of lucrative incentives. As a result, ransomware has recently dominated the cyberthreat landscape. Individuals, businesses, government agencies, universities, and hospitals, are targeted by ransomware attacks. For instance, in 2017, the Wannacry ransomware infected over 300,000 victims in 150 countries via the Shadow Brokers APT EternalBlue exploit. Petya ransomware was the first targeted ransomware attack, with most infections occurring in Ukraine. However, Petya has spread to over 60 countries. As a result, ransomware attacks continue to dominate the cyber security world, with an expected dramatic increase in targeted attacks. Due to the exponential growth of ransomware attacks, it is necessary to focus on this type of threat. Exploit kits, cryptocurrency, and ransomware-as-a-service (RaaS) are the primary factors accelerating the global crypto-ransomware outbreak. With RaaS, even inexperienced attackers can launch a crypto-ransomware attack against any organization [4]–[7].

The rest of the paper is organized as follows. Sections II, III and IV discussed this research's motivations, scope, and objectives. Finally, Sections V and VI discussed the research contributions and design. Section VII provided the literature review for this study. Section VIII described the dataset used in this work. Section IX described the framework design and development. Next, Section X described the testing and validation done in this study. Finally, Section I concluded the research and explained the possible future work for this research.

II. MOTIVATION

The crypto-ransomware attack is irreversible, and it is almost impossible to recover the files. Therefore, there is a need to identify it before it attacks the system and files. Crypto-

ransomware poses a significant threat, with new varieties and families being regularly discovered on the internet and the dark web. Furthermore, due to the encryption mechanisms utilized by these outbreaks, recovering from ransomware attacks is challenging [8]–[10]. In addition to the costs of downtime and the money that individuals and businesses may be compelled to pay as ransom, victims may suffer other consequences such as data loss, reputation loss, and even death [8], [11], [12].

III. RESEARCH SCOPE

This research is implemented only for the crypto-ransomware attack. However, this malware category is still persistent and creates massive damage in many crucial sectors [13], [14]. Furthermore, this research focused on crypto-ransomware targeting the Windows operating system since this platform is the most exciting target for the crypto-ransomware operators [15]–[17]. Besides that, most crypto-ransomware targets regular and average computer users, and most of them are running the Windows operating system [18]. The Windows operating system was chosen because it is the most frequently used platform in computer systems and is targeted by most ransomware attacks. The research will focus on crypto-ransomware that uses an encryption algorithm to encrypt its victim's data and files. Crypto-ransomware was chosen because the damage caused by this type of ransomware is often severe and irreversible [19], [20].

IV. RESEARCH OBJECTIVES

The first objective is to investigate ransomware behavior via Windows API calls. API is the set of instructions that every program uses to communicate with the operating system. Therefore, it is critical to analyze the ransomware's API to understand the ransomware's behavior better. The second goal is to develop an early detection framework for crypto-ransomware attacks. This is to mitigate the ransomware attack's damage. The third objective of this research is to create a dataset to identify crypto-ransomware in its early stages. This dataset contains critical data from the initial stages of a crypto-ransomware attack. The dataset will be used to train and test the machine learning classifier. Furthermore, this dataset can be used for future research in ransomware early detection.

V. RESEARCH CONTRIBUTIONS

In meeting the above objectives, this research has provided the following contributions. The first contribution is discovering important behaviors of crypto-ransomware attacks with the analysis of API produced. The second contribution is developing the RENTAKA framework to detect crypto-ransomware before triggering the mass unauthorized file encryption. The third contribution is an algorithm for determining the pre-encryption boundary and assists in extracting the required features. The fourth contribution is the crypto-ransomware early-stage behavior dataset that can aid future research using a machine learning approach. This research also filled the gap from previous research in focusing on the pre-encryption stage of the crypto-ransomware attack, which is a critical point; recovery is impossible after encryption happens. In addition, this research also provided a unique solution by combining the signature matching approach

and machine learning approach, which provided two levers of detection that can complement each other. Listed below are the contributions of this research:

- Identification of crypto-ransomware behavior during the early stages.
- Proposed a framework for crypto-ransomware early using a machine learning approach.
- Proposed an algorithm for pre-encryption features.
- Crypto-ransomware behavior dataset.

VI. RESEARCH DESIGN

The framework within which a researcher chooses the research methods and techniques is the research design. The design enables researchers to focus on appropriate research methods for the subject matter and establish a foundation for success in their studies. Therefore, this study is divided into four phases, as depicted in Fig. 1:

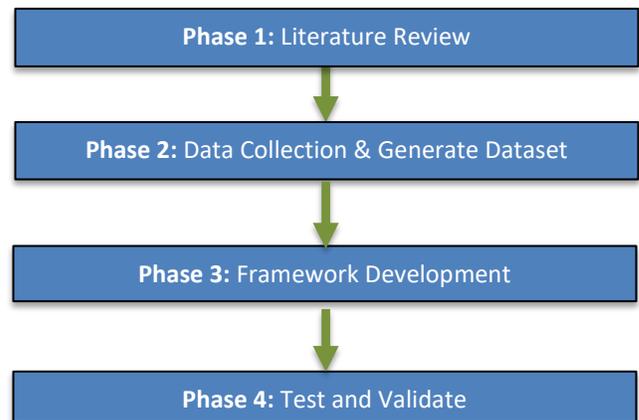


Fig. 1. Research Design.

VII. PHASE 1: LITERATURE REVIEW

A. Ransomware

Locker and crypto-ransomware are the two types of ransomware. The Locker ransomware merely affects the user interface, leaving the system and files intact. The Locker ransomware encrypts files and disables operating system features, including desktop apps and input/output utilities. Meanwhile, cryptographic ransomware, often known as crypto-ransomware, tries to extract money from victims by encrypting their files [21].

Crypto-ransomware encrypts user-related files using the cryptography features in the host operating system. The consequences of such ransomware are reversible only through the cryptographic keys possessed by a distant adversary, which sets it apart from other types of malware. Files that have been encrypted are renamed and given new extensions. Some of the most common ransomware encrypted file extensions are ".ccc", ".cerber", ".cerber2", ".cerber3", ".crypt", ".cryptolocker", ".cryptowall", ".ecc", ".ezz", ".locky", ".micro", ".zepto", and ".encrypted". It substitutes a fresh wallpaper with a ransom note for the original desktop background. Cryptolocker,

CryptoDefense, KeRanger, ZCryptor, Crysis, zCrypt, Locky, and WannaCry are just a few examples of crypto-ransomware [5], [8], [22].

The availability of development toolkits and the ease with which ransomware assaults can be traced from victims to attackers are the key factors driving the surge in ransomware attacks today. Before a ransomware attack can occur, it must first get access to the victim's computer [23]–[25].

Ransomware is commonly distributed using spear-phishing and exploit kits. Spear-phishing is a sophisticated email assault designed to deceive people or corporations into accessing a malicious website infected with malware. These emails frequently include attention-getting content from reputable sites to attract recipients to click on the offered link. Furthermore, it is common for an attacker to employ a series of commands or code to exploit the capabilities of a susceptible program. Finally, exploit kits, which can be used manually or automatically, assist hackers in finding flaws in software that would otherwise be impenetrable [26]–[28].

Ransomware has developed throughout time. Its many variations are being produced daily. As a result, there are a lot of ransomware families and their variants. Various obfuscation tactics are used for creating new versions, including garbage code insertion, variable renaming polymorphism, metamorphism, and packing [29].

Crypto-ransomware is malware that encrypts a victim's data and holds it hostage in exchange for money. Cybercriminals collect the ransom in the form of cryptocurrency, typically Bitcoins, to hide their identity. There are two types of ransomware: locker and crypto-ransomware. Crypto-ransomware is more common and offers a more significant threat than locker ransomware [30], [31].

WannaCry, Cryptolocker, Cryptowall, and Locky are examples of ransomware that have progressed from low-impact assaults like PC-Cyborg (also known as AIDS) to high-impact attacks like WannaCry, Cryptolocker, Cryptowall, and Locky. In addition, the number of ransomware variations has been rising since 2012. For example, ransomware variations increased from one to 193 between 2012 and 2016. As a result, ransomware became a significant threat to cybersecurity during this period. In addition, ransomware-as-a-Service (RaaS) families like Cryptolocker, CryptoWall, Locky, and TeslaCrypt also appeared in 2017, causing significant financial losses worldwide [32].

Crypto-ransomware assaults have become increasingly prevalent, allowing attackers to make millions of dollars per month. Around 180 million non-technical individuals were victimized by ransomware in 2017. In 2018, there were approximately 850 million ransomware assaults. In 2019 and 2020, ransomware is expected to have caused around \$11.5 billion and \$20 billion in global damage, respectively.

B. Crypto-ransomware Lifecycle

- **Deployment:** The crypto-ransomware must be able to install itself on the targeted system successfully. Phishing emails are the most typical way for ransomware to propagate. Cybercriminals use social

engineering techniques to persuade people to believe the email message and open the malicious file attached to it. Social engineering approaches include executables with appealing icons, Microsoft Office macros, and phishing files. Furthermore, ransomware spreads using malicious websites or exploit kits like Angler and Magnitude.

- **Installation:** The infection begins after a malicious payload successfully lands on the victim's platform. The malicious components are built using scripts, procedures, batch files, and other resources. Ransomware will make configuration changes to a Windows-based system, such as establishing unique registry keys in the registry to ensure harmful malware runs every time the computer reboots. Payload persistence, restricted system restoration, stealth mode, environment mapping, and privilege escalation are all features of more complex crypto-ransomware.
- **Command and Control:** After the ransomware has been installed, it begins interacting with its command and control server. This server provides ransomware with further instructions and a public encryption key. Next, the crypto-ransomware will try to connect to its C&C server, which the ransomware operator controls. Once the link has been established, it will provide information about the victim's computing platform and the encryption key.
- **Destruction:** The encryption stage begins after establishing effective contact with the targeted computer. The files are encrypted once the ransomware gets the encryption code and the location of the victim files. The encrypted files are renamed with a different extension when the original files are erased. Some ransomware variations add their name to any file as an extension. A list of the files that will be encrypted is included in the ransomware payload. Essential files like "WINDOWS," "Application Data," and "Temp" are excluded to keep the Windows operating system functioning. By erasing the volume shadow copies, ransomware prohibits the user from restoring them. Instead, it uses administrator rights to erase shadow copies of the Windows drive using the cmd.exe command.
- **Extortion:** The ultimate stage of a crypto-ransomware assault is extortion. Once the data have been fully encrypted, the next step is to inform them and persuade them to pay the suggested ransom. At this point, a windows pop-up or a desktop wallpaper with the ransom note appears on the screen. The directions on proceeding with the ransom payment are included in the ransom note. All ransomware has a different look and various texts in the ransom note. Finally, the crypto-ransomware final stage shows an extortion message demanding a ransom in exchange for the decryption key. Fig. 2 shows the steps in the crypto-ransomware lifecycle.



Fig. 2. Crypto-ransomware Lifecycle.

C. Early Detection

Crypto-ransomware is a type of malware that is relatively new. To our knowledge, only a few studies have been conducted on early detection. However, the emerging threat of crypto-ransomware piqued the interest of numerous researchers worldwide, who sought to develop a method for detecting it. Additionally, due to crypto-encryption ransomware's capability, this distinct characteristic can be used as a critical indicator for its early detection during the pre-encryption stage.

Crypto-ransomware pre-encryption detection detects it even before the encryption process begins. Due to the critical nature of detecting crypto-ransomware early in the attack lifecycle, several studies on pre-encryption detection of crypto-ransomware have been proposed.

It is more difficult to detect in the pre-encryption phase due to a lack of evidence that crypto-ransomware is present. Simultaneously, no unauthorised encryption activity occurs. The benefits of successfully detecting a crypto-ransomware infection at this level are that no files are lost and the ransomware is prevented from infecting additional hosts or networks.

In the case of crypto-ransomware, detecting it during the pre-encryption stage is very valuable. Due to the irreversible and irrecoverable nature of a crypto-ransomware attack, it is critical to detect it early, even before it begins encrypting the files. Several studies have proposed methods for detecting crypto-ransomware infections before encryption. Pre-encryption detection occurred before the start of file encryption activity. Detection is critical at this stage to prevent any files from being encrypted. The benefits of detecting crypto-ransomware activity at this level are incredibly beneficial for an organization's file, system, and network security. Apart from preventing any files from being encrypted, detection at this stage may alert system administrators to the infection as soon as possible, allowing security precautions to be taken in time.

Additionally, this proactive measure can help prevent the spread of crypto-ransomware to other endpoints or networks. Finally, detection enables system owners and administrators to

respond to an attack as soon as possible before significant damage is caused.

The basic steps in most crypto-ransomware lifecycle, as shown in Fig. 2 are further grouped into three sub-phases of attack: Pre-encryption, encryption, and Post-encryption. These sub-phases are depicted in Fig. 3.

- Pre-encryption – because any crypto-ransomware objective is to encrypt files in bulk, it is frequently designed to avoid detection by making a series of pre-attack API requests. Fig. 4 shows a list of activities during this stage.
- Encryption – at this attack level, unauthorized mass file encryption is taking place.
- Post-encryption – this is where the extortion takes place, by strategically notice the victim of the fate of the encrypted files and luring the system owner to execute the ransom payment.

Crypto-ransomware is a dreaded type of malware that has gained notoriety because of its fatal and irreversible effects on its victims. Due to the irreversible damage caused by ransomware, it is critical to notice these assaults quickly. The following is a list of the reasons why early detection of crypto-ransomware is critical:

- To avoid file loss and the need to pay the ransom (Kok et al., 2019).
- To detect ransomware attacks as early as possible to prevent data loss and stop ransomware self-propagation (Roy & Chen, 2019).
- Early detection can help users protect confidentiality and availability while limiting the probability of an attack and minimizing losses (Moussaileb, 2018).
- The value of detecting cryptographic ransomware is at the pre-encryption stage. It is useless after the encryption activity is completed because data loss has already happened.
- The damage done by crypto-ransomware is irreversible (Al-Rimy & Maarof, 2018).

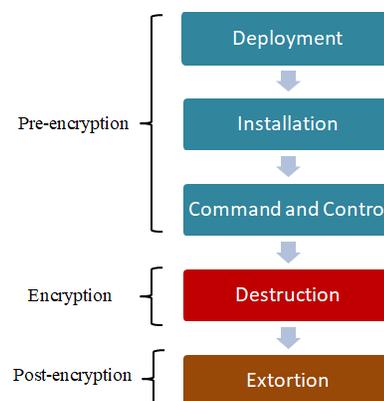


Fig. 3. Pre-encryption, Encryption and Post-encryption Stages.

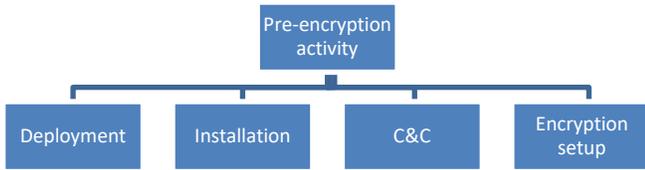


Fig. 4. Activities of a Crypto-ransomware during the Early Stage.

D. Related Work

Early detection and prediction are advantageous for varieties of malware where recovery is difficult and costly. Crypto-ransomware, for example, encrypts user files and withholds the decryption key until the perpetrators are paid a ransom. Unfortunately, as the frequency of crypto-ransomware attacks has grown in recent years, so has the research community’s attention to this issue. As a result, few studies examining various threat detection strategies have been conducted. Table I provided a summary of the related works.

EldeRan is a machine learning framework for detecting ransomware early in its lifecycle. As far as this research is concerned, it is the first of its kind in ransomware early detection. The framework looked at dynamic analysis data from ransomware samples. It also keeps track of events throughout the ransomware’s installation phase to capture ransomware features. As a result, EldeRan can operate without requiring advanced access to a ransomware family. The first restriction is that it is difficult to analyze and identify crypto-ransomware samples that have been silent for a long time or are waiting for a user-initiated trigger action [33].

As a ransomware early detection framework, the Pre-encryption Detection Algorithm, or PED A, was proposed [34]. The framework has two phases: PED A-Phase-I and PED A-Phase-2. API calls were collected after examining the samples for 30 seconds in the Cuckoo sandbox. PED A-Phase-I will use the learning algorithm (LA) to collect and analyze the Windows API calls generated by a suspicious sample. The LA can then assess whether the suspicious program was ransomware or not using API pattern recognition. This method ensures the most thorough identification of known and unknown ransomware, but it may lead to many false positives. PED A implemented a signature database for the samples and placed it in the Phase-II signature repository if the prediction was for ransomware.

Meanwhile, in PED A-Phase-II, the signature repository uses the signature matching method to detect ransomware at a far earlier level, namely the pre-execution step. Yet this approach only detects known ransomware, though it is proven accurate and quick despite its rigidity. PED A’s two phases resulted in two layers of early ransomware detection, guaranteeing that the victim’s data was not lost. This technique, however, was unable to detect ransomware that employed its encryption code and inherited the disadvantages of a signature-based approach [24].

For the early detection of crypto-ransomware, Alqah tani introduced the CRED framework [35]. Their study focuses on the flaws in currently existing ransomware early detection tools. They also presented a model capable of accurately

characterizing the attack lifecycle’s pre-encryption phase. This strategy is superior because it can overcome data insufficiency gathered during the pre-encryption phase by giving enough time before stopping data collection and using two categories of data: process-centric and data-centric. However, they did not present experimental data to indicate that their approach is superior to others because their study is preliminary.

A group of researchers used file system data to detect ransomware, including whether the contents appear to have been encrypted and the number of modifications made to the file type. As a result, the researchers recognized all 492 ransomware strains tested and prevented, with less than 33% of user data destroyed in each case [36].

TABLE I. DIFFERENCES OF CURRENT PRE-ENCRYPTION DETECTION FRAMEWORK

Framework	Elderan	PEDA	CRED
Pre-encryption boundary identification	Dynamic analysis runtime limited to first 20 seconds.	Identify first occurrence of CryptoAPI.	Using temporally correlated IRP-API based pre-encryption delineation method
Features	30967 features. (API, registry key operations, file operations, dropped files, embedded strings)	232 API calls	API calls and IRP
Dataset	RISS	RISS	Process-centric Data-centric

E. Challenges in Pre-encryption Detection

Crypto-ransomware operations are often disguised as legal user actions, mainly when crypto-ransomware does not require special rights and depends on cryptographic functionality similar to benign applications. Because most crypto-ransomware either implements cryptography or uses existing libraries, this is the case. Apart from that, all they have to do is read and write files.

Detecting ransomware is a race between the bad guys and the creators. New countermeasures push ransomware creators to improve their ransomware, resulting in new countermeasures. For ransomware scenarios, it may, for example, act more like legitimate software or a human user.

There are few strong evidence indicators during the early stages of a crypto-ransomware attack. For example, there is no evidence that many files were encrypted during these early stages. Furthermore, no encryption action is taking place. Strange file extensions, unauthorized changes to the desktop wallpaper, the appearance of a ransom letter, increased CPU utilization, or system slowdown are not visible symptoms of ransomware infection. As a result, the evidence available during the early stages of the investigation is inadequate to evaluate whether the described behaviors are crypto-ransomware-like. It’s impossible to tell whether the listed current activity belongs to a benign program or a crypto-ransomware because there was no significant unintended encryption activity during the early phases [20], [34], [37].

There are few significant indicators during the early phases of a crypto-ransomware assault. The fundamental reason is that

there is no indication of illicit file encryption during the early stages [38], [39]. Furthermore, there is no encryption operation in progress. Strange file extensions, unauthorized desktop wallpaper changes, the appearance of a ransom letter, increased CPU use, and system slowdown are not indicative of a ransomware infestation. As a result, the information supplied at the investigation's outset is insufficient to determine whether the described actions are crypto-ransomware-like. Determine whether the related current activity is owned by benign software or crypto-ransomware due to the lack of significant illegal encryption activity in the early stages. The data on the victim's PC is encrypted using a robust encryption method in any crypto-ransomware attack.

VIII. PHASE 2: DATASET

The dataset was from the Resilient Information System Security (RISS) research group from Imperial College London in 2016. This dataset was selected because it has API data for ten ransomware families and a good selection of goodware. The dataset was created using a dynamic analysis approach for 582 samples of ransomware and 942 samples of benign program. The data are captured in five main categories with 30,067 features. API calls have 232 features. Two groups of researchers used this dataset for works on crypto-ransomware early detection frameworks [33], [34].

As far as this research is concerned, the dataset on crypto-ransomware behavior is still lacking. However, the RISS dataset is by far the best dataset available for ransomware behavior, and this is shown by the works done by Elderan and PEDDA [33], [34]. Tables II and III provided some information about the RISS dataset used in this study.

TABLE II. RANSOMWARE FAMILIES IN RISS DATASET

No.	Sample name	Count
1	Critroni	50
2	Cryptlocker	107
3	Cryptowall	46
4	Kollah	25
5	Kovter	64
6	Locker	97
7	Matsnu	59
8	Pgpocoder	4
9	Reveton	90
10	Teslacrypt	6
11	Trojan-ransom	34

TABLE III. CATEGORIES OF DATA IN RISS DATASET

Category	Count
API	232
Registration key	346
Dropped file	6622
Files and directory operation	7500
Embedded string	16267
Total	30967

These researchers successfully used the RISS dataset from different institutions and produced acceptable results. Another dataset is from The Zoo malware repository, which provides ransomware binaries that can be downloaded and analyzed into dynamic analysis sandboxes such as Cuckoo Sandbox.

IX. PHASE 3: FRAMEWORK DEVELOPMENT

Given the size and variety of threats we face today, having solutions to detect unknown crypto-ransomware attacks before unauthorized mass file encryption takes place seems necessary. In addition, it is essential to protect user data from any variants of crypto-ransomware attacks with zero data loss.

Monitoring API calls made by crypto-ransomware makes it possible to design an early detection framework to halt crypto-ransomware attacks, including those using sophisticated encryption capabilities.

We proposed a pre-encryption detection framework for crypto-ransomware using a machine learning approach, RENTAKA, to protect user data from being encrypted. The framework is depicted in Fig. 5. Based on detailed investigations of most cases, ransomware-specific events and processes are heavily related to Application Programming Interface (API) calls for the Windows platform. User-level malware like ransomware requires the invocation of system calls to interact with the operating system (OS) to execute its malicious actions. Application Programming Interface (API) calls are the functions that a program utilizes in its execution. In other words, API calls are a set of routines provided by the OS for building applications in which each API call performs a specific task. The API calls is extracted through dynamic analysis after executing the ransomware sample in a sandbox environment. We demonstrate that our proposed solution can detect crypto-ransomware in the pre-attack stage and achieve zero data loss against current ransomware families. Furthermore, as shown in Table IV, we also proposed an algorithm to extract the data related to the pre-encryption stages.

TABLE IV. PSEUDOCODE FOR PRE-ENCRYPTION BOUNDARY ALGORITHM

1. Sample executes in sandbox
2. Run dynamic analysis
3. Extract the behavioral log
4. Locate APIstat cluster
5. Find encryption API
 - a. If found encryption API, flag it as "ENC"
 - b. Extract all API before the ENC flag
 - c. Store in a file
 - d. Kill sample execution, repeat with next sample

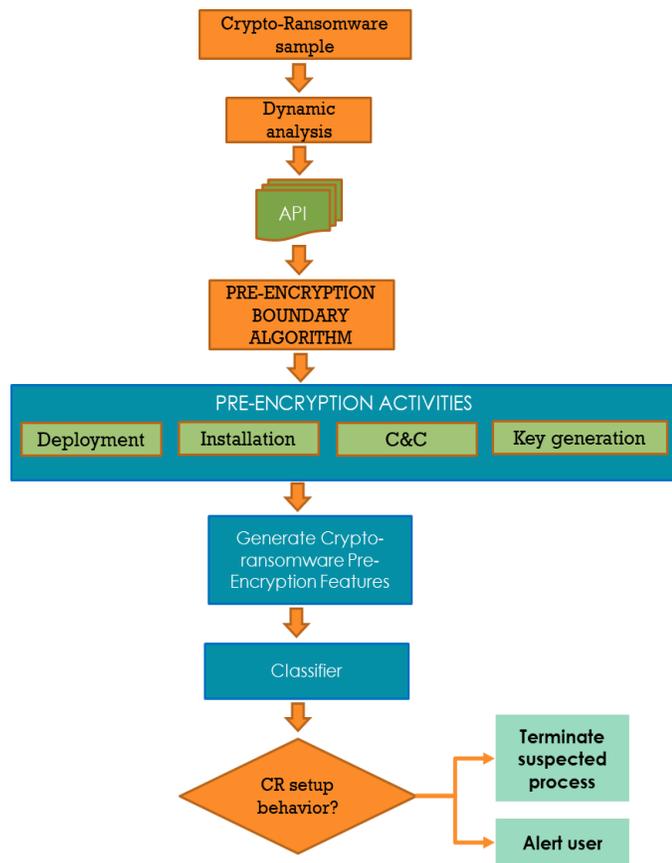


Fig. 5. The Proposed RENTAKA Framework.

X. PHASE 4: TEST AND VALIDATE

The proposed model is tested and validated on a real-world corpus of ransomware samples. The results show that API call features accurately distinguish between ransomware binaries and benign ones. Furthermore, the relevant feature selection process can improve the model building time without compromising the accuracy of the malware detection system.

This study experimented with 80 features using five different classification algorithms: Random Forest, Naïve Bayes, SVM, kNN, and J48. Based on our experiments, support vector machines (SVMs) performed with the best accuracy and TPR, 97.05% and 0.995, respectively. The second-best result is the Random Forest classifier, with 96.39% accuracy. Finally, J48 performs with the lowest accuracy, which is 94.75%. The overall results from our experiments are listed in Table V.

TABLE V. RESULTS FROM MACHINE LEARNING CLASSIFIERS

Classifier	Accuracy	TPR	FPR
Random Forest	96.3934%	0.984	0.071
Naïve Bayes	80.9836%	0.781	0.142
SVM	97.0492%	0.995	0.071
kNN	96.0656%	0.979	0.071
J48	94.7541%	0.979	0.106

XI. CONCLUSION AND FUTURE WORK

This paper discussed the ransomware categories, attack lifecycle, analysis approaches, detection techniques, and related works in its detection. This paper also provided the challenges of crypto-ransomware early detection. We proposed a ransomware detection scheme using a machine learning classifier. Based on our experiments, support vector machine (SVM), one of the supervised machine learning algorithms, performed the best accuracy and TPR. Crypto-ransomware attacks are very dynamic, and it is moving toward becoming a kind of targeted attack. Therefore, early detection systems with machine learning-based classification algorithms are needed to mitigate crypto-ransomware attacks. For future work, we will test with more extensive samples and improve the pre-encryption boundary algorithm. The encryption boundary identification algorithm is a crucial part of this research. It defines the number of features to be used for building the machine learning model.

REFERENCES

- [1] S. Bistarelli, M. Parrocchini, and F. Santini, "Visualising bitcoin flows of ransomware: WannaCry one week later," CEUR Workshop Proc., vol. 2058, pp. 1–8, 2018.
- [2] S. H. Kok, A. Abdullah, N. Z. Jhanjhi, and M. Supramaniam, "Ransomware, Threat and Detection Techniques: A Review," IJCSNS Int. J. Comput. Sci. Netw. Secur., vol. 19, no. 2, pp. 136–146, 2019.
- [3] T. Dargahi, A. Dehghantanha, P. N. Bahrami, M. Conti, G. Bianchi, and L. Benedetto, "A Cyber-Kill-Chain based taxonomy of crypto-ransomware features," J. Comput. Virol. Hacking Tech., vol. 15, no. 4, pp. 277–305, 2019, doi: 10.1007/s11416-019-00338-7.
- [4] A. Zimba and M. Chishimba, "On the Economic Impact of Crypto-ransomware Attacks: The State of the Art on Enterprise Systems," Eur. J. Secur. Res., vol. 4, no. 1, pp. 3–31, 2019, doi: 10.1007/s41125-019-00039-8.
- [5] M. Akbanov, V. G. Vassilakis, and M. D. Logothetis, "WannaCry ransomware: Analysis of infection, persistence, recovery prevention and propagation mechanisms," J. Telecommun. Inf. Technol., no. 1, pp. 113–124, 2019, doi: 10.26636/jtit.2019.130218.
- [6] J. A. H. Silva, L. Isabel, and B. López, "A Survey on Situational Awareness of Ransomware Attacks — Detection and Prevention Parameters," 2019, doi: 10.3390/rs11101168.
- [7] A. O. Almashhadani, M. Kaijali, S. Sezer, and P. O’Kane, "A Multi-Classifer Network-Based Crypto Ransomware Detection System: A Case Study of Locky Ransomware," IEEE Access, vol. 7, pp. 47053–47067, 2019, doi: 10.1109/ACCESS.2019.2907485.
- [8] B. A. S. Al-rimy, M. A. Maarof, and S. Z. M. Shaid, "Ransomware threat success factors, taxonomy, and countermeasures: A survey and research directions," Comput. Secur., 2018, doi: 10.1016/j.cose.2018.01.001.
- [9] A. Alqahtani and F. T. Sheldon, "A Survey of Crypto Ransomware Attack Detection Methodologies: An Evolving Outlook," pp. 1–19, 2022.
- [10] M. Rhode, P. Burnap, and K. Jones, "Distillation for run-time malware process detection and automated process killing," pp. 1–12, 2019, [Online]. Available: <http://arxiv.org/abs/1902.02598>.
- [11] M. Wojnowicz, G. Chisholm, B. Wallace, M. Wolff, X. Zhao, and J. Luan, "SUSPEND: Determining software suspiciousness by non-stationary time series modeling of entropy signals," Expert Syst. Appl., vol. 71, no. March 2017, pp. 301–318, 2017, doi: 10.1016/j.eswa.2016.11.027.
- [12] Z. A. Genç, G. Lenzini, and P. Y. A. Ryan, "The Cipher, the Random and the Ransom: A Survey on Current and Future Ransomware," 2017, [Online]. Available: <http://orbilu.uni.lu/bitstream/10993/32574/1/GLR2017.pdf>.

- [13] S. Homayoun, A. Dehghantaha, M. Ahmadzadeh, S. Hashemi, and R. Khayami, "Know Abnormal, Find Evil: Frequent Pattern Mining for Ransomware Threat Hunting and Intelligence," *IEEE Trans. Emerg. Top. Comput.*, vol. 6750, no. c, pp. 1–1, 2017, doi: 10.1109/TETC.2017.2756908.
- [14] S. Homayoun et al., "DRTHIS: Deep ransomware threat hunting and intelligence system at the fog layer," *Futur. Gener. Comput. Syst.*, vol. 90, pp. 94–104, 2019, doi: 10.1016/j.future.2018.07.045.
- [15] N. Hampton, Z. Baig, and S. Zeadally, "Ransomware behavioural analysis on windows platforms," *J. Inf. Secur. Appl.*, vol. 40, pp. 44–51, 2018, doi: 10.1016/j.jisa.2018.02.008.
- [16] A. K. Maurya, N. Kumar, A. Agrawal, and R. A. Khan, "Ransomware : Evolution , Target and Safety Measures," *Int. J. Comput. Sci. Eng. Open Access Res. Pap.*, no. 1, pp. 80–85, 2018, [Online]. Available: http://www.ijcseonline.org/pub_paper/12-IJCSE-02742.pdf.
- [17] A. Tandon and A. Nayyar, *A Comprehensive Survey on Ransomware Attack: A Growing Havoc Cyberthreat*, vol. 2, no. Proceedings of ICDMAI 2018. Springer Singapore, 2019.
- [18] M. A. Salah, M. Fadzli Marhusin, and R. Sulaiman, "Malware Research Directions: A Look into Ransomware," *Asian Journal of Information Technology*, vol. 16, no. 6, pp. 458–464, 2017.
- [19] S. H. Kok, A. Abdullah, and N. Z. Jhanjhi, "Early Detection of Crypto-Ransomware using Pre-Encryption Detection Algorithm," *J. King Saud Univ. - Comput. Inf. Sci.*, no. xxxx, 2020, doi: 10.1016/j.jksuci.2020.06.012.
- [20] B. A. S. Al-rimy and M. A. Maarof, "A 0-Day Aware Crypto-Ransomware Early Behavioral Detection Framework," 2018, doi: 10.1007/978-3-319-59427-9.
- [21] K. Savage, P. Coogan, and H. Lau, "The Evolution of Ransomware," *Secur. Response*, p. 57, 2015, doi: 10.5437/08953608X5403011.
- [22] K. Cabaj, M. Gregorczyk, and W. Mazurczyk, "Software-Defined Networking-based Crypto Ransomware Detection Using HTTP Traffic Characteristics," 2015, [Online]. Available: <https://arxiv.org/ftp/arxiv/papers/1611/1611.08294.pdf>.
- [23] V. C. Craciun, A. Mogage, and E. Simion, "Trends in design of ransomware viruses," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 11359 LNCS, pp. 259–272, 2019, doi: 10.1007/978-3-030-12942-2_20.
- [24] U. Urooj, M. Aizaini Bin Maarof, and B. Ali Saleh Al-Rimy, "A proposed Adaptive Pre-Encryption Crypto-Ransomware Early Detection Model," 2021 3rd Int. Cyber Resil. Conf. CRC 2021, pp. 7–12, 2021, doi: 10.1109/CRC50527.2021.9392548.
- [25] D. Y. Kao, S. C. Hsiao, and R. Tso, "Analyzing WannaCry Ransomware Considering the Weapons and Exploits," *Int. Conf. Adv. Commun. Technol. ICACT*, vol. 2019-Febru, no. 2, pp. 1098–1107, 2019, doi: 10.23919/ICACTION.2019.8702049.
- [26] J. Kaur, F. Jaafar, and P. Zavorsky, *An Empirical Analysis of Crypto-Ransomware Behavior*. 2018.
- [27] S. Chadha, "Ransomware : Let ' s Fight Back !," pp. 925–930, 2017.
- [28] R. Moussaileb, N. Cuppens, J. L. Lanet, and H. Le Bouder, "Ransomware Network Traffic Analysis for Pre-encryption Alert," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 12056 LNCS, pp. 20–38, 2020, doi: 10.1007/978-3-030-45371-8_2.
- [29] Q. Chen and R. A. Bridges, "Automated behavioral analysis of malware: A case study of wannacry ransomware," *Proc. - 16th IEEE Int. Conf. Mach. Learn. Appl. ICMLA 2017*, vol. 2017-Decem, pp. 454–460, 2017, doi: 10.1109/ICMLA.2017.0-119.
- [30] K. Liao, Z. Zhao, A. Doupe, and G. J. Ahn, "Behind closed doors: Measurement and analysis of CryptoLocker ransoms in Bitcoin," *eCrime Res. Summit, eCrime*, vol. 2016-June, pp. 1–13, 2016, doi: 10.1109/ECRIME.2016.7487938.
- [31] M. S. Rosli et al., "Ransomware Behavior Attack Construction via Graph Theory Approach," vol. 11, no. 2, pp. 487–496, 2020.
- [32] B. N. Giri and N. Jyoti, "The Emergence of Ransomware," doi: 10.1177/0306396801432003.
- [33] D. Sgandurra, L. Muñoz-González, R. Mohsen, and E. C. Lupu, "Automated Dynamic Analysis of Ransomware: Benefits, Limitations and use for Detection," 2016, doi: 10.15199/48.2015.11.48.
- [34] S. Kok, A. Abdullah, N. Jhanjhi, and M. Supramaniam, "Prevention of Crypto-Ransomware Using a Pre-Encryption Detection Algorithm," *Computers*, vol. 8, no. 4, p. 79, Nov. 2019, doi: 10.3390/computers8040079.
- [35] A. Alqahtani, M. Gazzan, and F. T. Sheldon, "A proposed Crypto-Ransomware Early Detection(CRED) Model using an Integrated Deep Learning and Vector Space Model Approach," 2020 10th Annu. Comput. Commun. Work. Conf. CCWC 2020, pp. 275–279, 2020, doi: 10.1109/CCWC47524.2020.9031182.
- [36] N. Scaife, H. Carter, P. Traynor, and K. R. B. Butler, "CryptoLock (and Drop It): Stopping Ransomware Attacks on User Data," *Proc. - Int. Conf. Distrib. Comput. Syst.*, vol. 2016-Augus, pp. 303–312, 2016, doi: 10.1109/ICDCS.2016.46.
- [37] M. Rhode, P. Burnap, and K. Jones, "Early Stage Malware Prediction Using Recurrent Neural Networks," 2017, doi: 10.1016/j.cose.2018.05.010.
- [38] A. Kharraz, W. Robertson, D. Balzarotti, L. Bilge, and E. Kirda, "Cutting the gordian knot: A look under the hood of ransomware attacks," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 9148, pp. 3–24, 2015, doi: 10.1007/978-3-319-20550-2_1.
- [39] M. Patyal, S. Sampalli, Q. Ye, and M. Rahman, "Multi-layered defense architecture against ransomware.," *Int. J. Bus. Cyber Secur.*, vol. 1, no. 2, pp. 52–64, 2017, [Online]. Available: <http://ezproxy.umuc.edu/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=bth&AN=121205538&site=eds-live&scope=site>.

Non-contact Facial based Vital Sign Estimation using Convolutional Neural Network Approach

Nor Surayahani Suriani, Nur Syahida Shahdan, Nan Md. Sahar, Nik Shahidah Afifi Md. Taujuddin
Dept. of Electronics Engineering, Universiti Tun Hussein Onn Malaysia
Batu Pahat, Johor, Malaysia

Abstract—A rapid heart rate may indicate early diagnosis of heart disease, which could result in abrupt mortality if a heart attack occurs while exercising. A fatal incident is usually precipitated by a heart attack while strenuously exercising. This paper proposed invasive health monitoring through remote photoplethysmography (rPPG) analysis captured by RGB video camera to measure a wide range of biological data. A non-contact facial-based vital signs prediction can facilitate checking pulse rate and respiration rate regularly. Several studies have been conducted on evaluating rPPG signals under a variety of static conditions and little head movement, including different skin tones, angles of the camera, and distance from the camera. A study of heart rate (HR) and breathing rate (BR) data from facial videos for fitness applications were presented in this paper. Most studies still do not have a way to measure vital sign estimation especially for physical activity application from facial videos. The face detector was applied based on three regions of interest on facial landmarks for vital sign estimation. Then, the rPPG method with convolutional neural network (CNN) is presented to construct a spatio-temporal mapping of essential characteristics and estimate the vital sign from a sequence of facial images of people after doing various types of exercises. This will allow people to keep track of their health while exercising and creating a tailored training program based on their physiological preferences. The absolute error (AE) between the estimated HR and the reference HR from all experiments is 2.16 ± 2.2 beats/min. While the AE for the estimated BR from the references BR are 1.53 ± 2.3 beats/min.

Keywords—rPPG; remote heart rate estimation; respiration rate; fitness

I. INTRODUCTION

HR and BR are key health indicators for monitoring heart and lung function, especially during this critical COVID-19 pandemic period. Monitoring post-COVID-19 patients' physical activity responses helps in the recovery process. To monitor vital indicators including respiration rate, oxygen saturation, and heart rate, contactless measurement is critical. Infections, fever, asthma, and breathing issues are all signs of a higher HR. Pneumonia or lung illness might create an unstable BR. BR can be influenced by a variety of circumstances, including exercise, emotions, and injuries. Shortness of breath and cough was noted by nearly 10% and 25% of COVID-19 patients, respectively [1].

Our research goal is to assess the efficacy of our prototype of health monitoring status, which will include a built-in camera for non-contact vital sign estimation. The aim of this study presented in this article shows the potential of

estimating HR and BR from facial video streams. We also intended to validate the performance of our prototype so that it could be integrated with the CNN model as a comprehensive non-invasive system to aid in the creation of our healthcare system. The proposed work demonstrates the efficacy of a CNN-based model for post-exercise datasets. The dataset was created to expand the application of our prototype health monitoring system.

Remote photoplethysmography (rPPG) approaches use a webcam or a mobile camera to determine an individual's heartbeat from pixel variation in human skin surface induced by cardiac activity. The principle that blood absorbs more light than other surrounding tissue is used to investigate variations in blood volume transmission on the skin surface using an optics-based rPPG approach. In general, the rPPG procedure entails detecting and tracking the individuals' skin colour changes. Heart rate, heart rate variability, and respiration rate were evaluated using the tracking signals.

Several studies have reported that the potential of rPPG in heart rate estimation is quite promising. Reflected light and movement interferences, however, continue to cause significant challenges. Furthermore, rather than BR, rPPG signals are frequently applied to measure HR. This is because rPPG's frequency properties may not be reliable in estimating respiratory rate. Traditional rPPG approaches have some drawbacks because they rely on assumptions like simplified noise reduction and a skin optical reflection model. Furthermore, remote rPPG approaches will produce unstable results in real-world circumstances where patients are in motion. Deep learning breakthroughs have considerably enhanced the performance of rPPG approaches [2].

This paper describes how to use spatial-temporal facial features mapping to train the CNN model and estimate the rPPG matrix vector on the network's final layers. The single vector represents the rPPG signal, which will be processed further to estimate the HR and BR. Compared to previous work [11-13] and [16], our method addresses the efficacy of vital sign estimation from facial video for fitness or involving physical activity. Our proposed method also addresses the challenges to improve the system robustness. As a result, the following are the paper's contributions:

1) Limit the facial regions to only the forehead, cheeks, and nose to form spatial-temporal features for the CNN model to increase the reliable information for blood variation on face. Our approach improves the signal-to-noise ratio (SNR) for a better signal quality assessment of physiological signals.

2) Calculate HR and BR from the estimated rPPG signal, which is the CNN output. Our approach is in line with deep learning research community where many CNN frameworks have been successfully developed for detection and estimation. Further signal processing algorithms are proposed to estimate the HR and BR from the estimated rPPG signals.

3) For post-exercises with more variable heart rates, we use the rPPG signal to estimate HR and BR using CNN. From the best of our knowledge, very little study includes non-contact facial of different conditions performing several physical exercises especially for fitness development application.

The remainder of this paper is structured as follows. Section 2 provides a synopsis of related work areas. Section 3 describes how to use CNN architecture to effectively estimate rPPG and measure HR and BR after subjects perform several physical exercises. Section 4 goes over the experimental setup and metrics for evaluating performance. Section 5 summarises the findings and discusses potential future work.

II. RELATED WORK

Previously, pulse oximetry was used to obtain non-invasive vital sign estimates by evaluating the PPG signal at different wavelengths. While non-invasive rPPG signal monitoring with a video camera is a viable technique for vital signs monitoring without even any electrodes or sensors directly contacting patients. The most of of rPPG estimation research was conducted under different illumination conditions, head movements, colour skin variations, and camera distance. All of these constraints were established using traditional methods. Deep learning has recently received a lot of attention.

A. Conventional Methods

The rPPG's early work is entirely based on a signal processing approach. To generate colour signals, skin detection is applied to a selected RoI and the average is tracked over time. The green channel of rPPG signals produced the strongest signal, which roughly corresponds to an oxyhemoglobin absorption peak. After that, the Blind Source Separation (BSS) technique is used to separate the rPPG signals from the noise [3].

The HR could be extracted using the Fast Fourier Transform (FFT), which also indicates the respiratory rate based on the RoI of the entire face. The independent component analysis (ICA) method predicts HR based on the largest spectral peak between 0.75 and 4Hz. Poh et al. [4] used viola-jones face detection to determine the mean intensity of the red (R), green (G), and blue (B) colour channels. Then, ICA was used to demix the pulse signal from the raw RGB signals. To handle skin-tone mismatch and subject motions, combined RGB channels must be normalized. Cheng et al. [5] use independent vector analysis to separate the facial components from the background regions in order to address illumination artefacts. Wang et al. [6] proposed modified projection orthogonal to the skin tone to extract pulse. Haan et al. [7] proposed CHROM method where the RGB channels were projected into the chrominance subspace to eliminate

motion components. Chen Lin et al. [8] proposed motion index method to develop a real-time contactless pulse rate status monitoring of head motion modelling using trajectories of tracked feature points. The BR in [9] was calculated using heart rate variability (HRV). Band-pass filtering and spectral analysis can be used to extract changes in the rPPG waveform. Similarly, in [10], the FFT is used to estimate HR from the rPPG signal, and motion analysis is used to estimate BR.

B. Deep Learning Methods

HR convolutional neural network (HR-CNN) detects regions of interest (RoI) in a pretrained Convolutional Neural Network (CNN) model, extracts de-noised signals, and predicts HR using an estimator. Spetlik et al. [11] proposed end-to-end HR estimation with a single scalar value of predicted HR as the output network. Qui et al. [12] created an EVM-CNN architecture that includes face detection, feature extraction, and estimation. The RoI defined the central part of the human face and formed spatial decomposition and temporal filtering based on 68 landmarks. The CNN was used to estimate HR based on these feature images. Chen and Mcduff [13] addressed subject motion issues using DeepPhys, a deep convolutional network. To estimate the HR signal, the network learns spatial masks and extracts features from blood volume pulse (BVP) data. Yu et al. [8] used deep spatiotemporal networks to reconstruct rPPG signals from raw facial videos as well. The measured rPPG signal has peaks that correspond to the R peak of ground truth electrocardiogram (ECG) signals. The other CNN model for HR estimation developed in [14] is trained using transfer learning on images constructed from synthetic rPPG signals. The synthetic rPPG signals are generated by interpolating BVP or ECG signals.

Another paper in [15] developed and trained a 2D CNN for skin segmentation on both skin and non-skin region samples. The detected skin region was then subjected to conventional rPPG algorithms (ICA and PCA) for HR estimation. Furthermore, the RoI colour information was extracted using the Generative Adversarial Network (GAN) architecture [16]. This method was used to create a high-quality noiseless rPPG signal in order to improve HR accuracy performance.

The research in [17] combines 2D CNN with Residual Neural Network (RNN) models such as Long Short-Term Memory (LSTM) to compare the performance with other HR-CNN models. The facial input was fed into a 2D CNN, which then extracted spatial features from RGB frames of video. In the context of temporal domain, RNN was used to form spatial features. Other than that, the Siamese-rPPG algorithm is based on Siamese 3D CNN [18]. The proposed framework is intended to overcome a variety of noises on various facial appearances. For a better pulse pixels extraction, the forehead and cheek regions were chosen as the RoI to extract significant rPPG information. The predicted rPPG signals were produced by fusion branches at the intermediate layer with two 1D convolutional layers, and average pooling layer. Table I shows the summary of related works using CNN for HR estimation based on facial videos. Most of the outcome was tested on public standard datasets like PURE, MAHNOB and COHFACE.

TABLE I. SUMMARY OF RELATED WORKS USING CNN

Ref.	Methods	MAE	RMSE
11	HR-CNN	10.09	13.14
12	EVM-CNN	6.85	6.95
13	DeepPhysPCA	2.35	4.50
16	Deep-HR	3.41	0.027

Instead of proposing a new framework to improve estimation performance through the use of deep learning-based methods, more understanding of CNN-based methods is required to clarify how it works with rPPG technology. Some research has also been conducted on the constraint and sensitivity of CNN-based networks in rPPG technology. According to the paper in [19], CNN for rPPG signal extraction is a learning-based information related to PPG signals, and the training is easily affected by the delay between the video data and the ground truth data. The CNN-based methods have some limitations, such as a limited number of frames. Some frameworks are not ideal for long-term signal estimation, possibly because they are only supporting for a specific dataset.

This paper addresses HR and BR estimation using the CNN method for subjects after they have completed a series of physical exercises in order to gain more valuable insights into the effectiveness of deep learning approaches in HR and BR prediction. Based on region-based skin detection, the recorded facial videos will be fed into CNN to create a spatio-temporal map of skin region. HR and BR are expected to be estimated using the spatiotemporal map. The comparison is made with ground truth data collected by contact-based HR and BR monitoring as presented in [20].

III. METHODOLOGY

This section will go into detail about several steps, such as preprocessing facial videos to extract skin regions and training spatio-temporal networks for rPPG estimation. Further investigation is required to calculate the HR and BR from our recorded dataset.

A. Preprocessing

The steps for preprocessing are as shown in Fig. 1. First, we convert the RGB input image to grayscale colour. Then, using Haar cascade for face detection, find the RoI in the input videos. The RoI of the face is divided into four sections: the Forehead, Eyes, Cheeks (including the Nose and Mouth), and Chin. The splitting steps are used to obtain local face features which capable to form sufficient RoI within the face area [22]. In Fig. 2, only two selected regions, the Forehead (RoI1) and both Cheeks and Nose (RoI2), serve as image training sequences for the CNN model. According to [23, 24], these areas typically contain more BVP information with the highest absorption region and are less affected by non-rigid motion such as smiling or eye blinking.

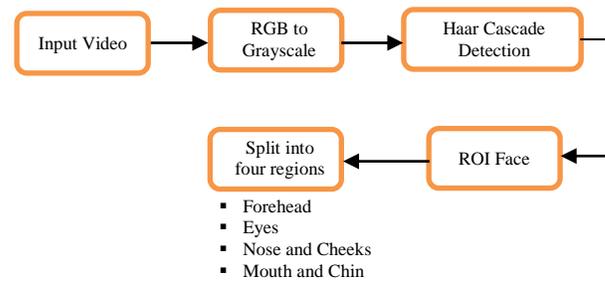


Fig. 1. Region based Skin Detection on Forehead, Cheeks, and Nose Areas.

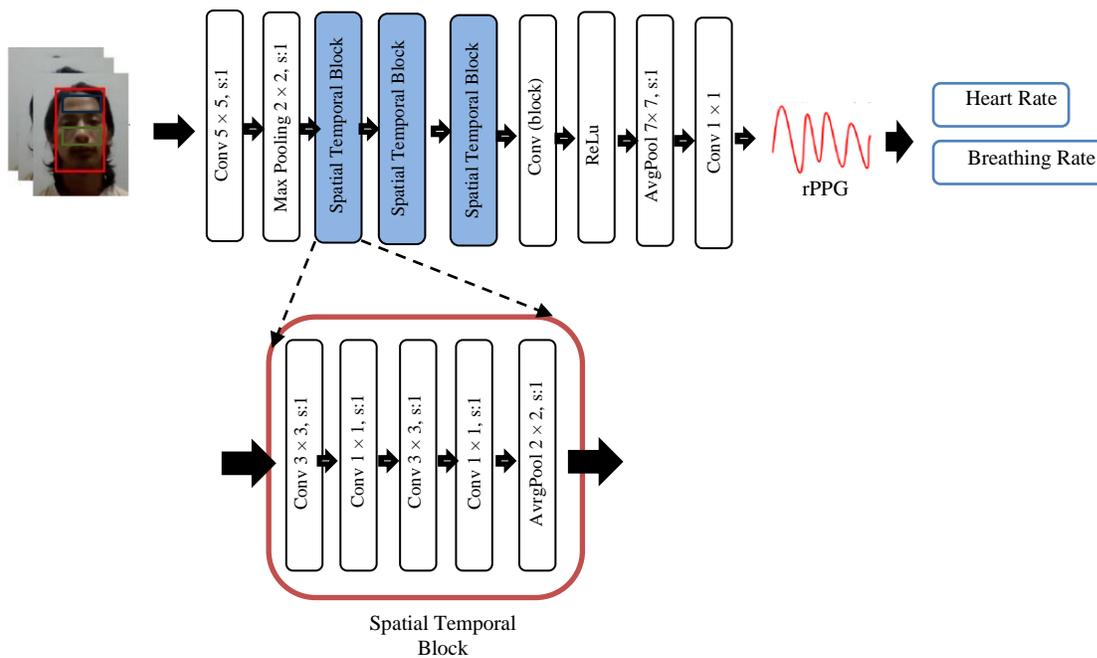


Fig. 2. The Overall Architecture from Region-based Skin Detection as an Input to the CNN and Spatio Temporal Convolutional Blocks. The Output is Single Scalar Vector of Predicted rPPG Signal.

B. Convolutional Neural Network

CNN was used specifically used to extract and estimate the single scalar vector of HR prediction. The CNN network is made up of multiple layers. Table II depicts the nine-layer CNN model structure. The first layer is a convolution kernel with a size of 5×5 and a stride of 1, which produces 64 feature maps. To reduce noise in extraction, the pool layer of the second layer used to perform maximum down-sampling of 2×2 . The next three layers are a spatio-temporal convolutional block with a 10×10 kernel in the convolution layer and a 2×2 kernel in the max-pooling layer. The high-level image features are generated by 200 feature maps. The next convolution layer has 500 feature maps, a kernel size of 4, and an average pooling layer with a 7×7 kernel. Convolution process, stride size, and pooling filter are used to minimise information loss while maintaining accurate physiological features for rPPG estimation.

The predicted rPPG signals are indicated by the output of spatio-temporal CNN from a single scalar vector. Each sample image was normalised to a size of 64×64 , with additional padding. Padding is required for all convolutions to maintain a consistent size. The non-activation ReLU function was used with a learning rate of 0.0005, a batch size of 200, and a training time period of 30.

C. Interbeat Interval

As shown in Fig. 3, peak detection is used to locate individual beats in the extracted rPPG signal. As a result, the inter-beat-interval (IBIs) is extracted from the rPPG signal. The IBIs are correspond to the time intervals between consecutive beats. Then, filters applied to the extracted IBIs to remove false positive/negative peak detections.

TABLE II. NETWORK ARCHITECTURE

CNN Layer	Layer Type	No. of Feature Map	Input Size
Layer 1	Conv.	20	64 x 64
Layer 2	mPool	20	32 x 32
Layer 3	ST Conv Block	200	16 x 16
Layer 4	ST Conv Block	200	8 x 8
Layer 5	ST Conv Block	200	8 x 8
Layer 6	ST Conv Block	200	4 x 4
Layer 7	ReLU	500	4 x 4
Layer 8	AvrgPool	16	1 x 1
Layer 9	Conv.	1	1 x 1

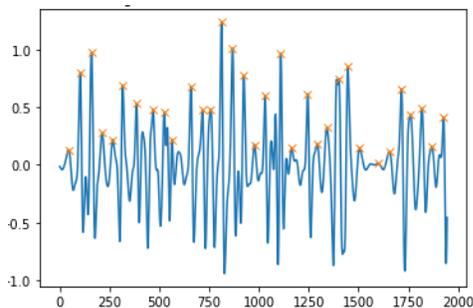


Fig. 3. Peak Detection of rPPG Signal.

D. Heart Rate Calculation

For valid peaks, the absolute IBI sequential difference should be less than 0.5 seconds. The HR is calculated by averaging all IBI over a time window and computing the inverse signals [21]. The IBI series is calculated as $IBI_{t_n} = t_n - t_{n-1}$ where t_n is the time of n -th detected peak. To simplify, $HR = 1/\overline{IBI}$ where \overline{IBI} is the mean of all inter-beat intervals within the time window. So, multiplying the HR in beats-per-minute by 60 yields the HR in beats-per-minute. For further analysis, the heart rate data is saved in a .csv file.

E. Respiration Rate Calculation

The normal respiration rate while people at resting condition is 12 to 20 beats per minute (bpm) [25]. Less than and more than this range are considered unstable respiration rates, indicating a slower (13bpm) or faster (>20bpm) breathing rate. Peak detection of the rPPG signal was used to estimate the BR using a time-domain technique. The duration of BR is defined as the time lag between the first and last detected peak. As a result, using spectral analysis on the estimated HR signal, the BR can be estimated. The power spectral density (PSD) was calculated, and the highest frequency amplitude within a plausible frequency range was chosen to represent the respiratory signal. The plausible respiratory frequency range was set from 0.1Hz to 0.4Hz.

F. Performance Metric Evaluation

The metrics evaluation used root mean square error (RMSE) and mean absolute error (MAE) to quantify the performance of proposed methods between predicted HR and BR rate and the ground truth. The fit standard error, or RMSE, was used to evaluate the best fitting of both estimated and ground truth data. As a result, as the RMSE decreases, so does the goodness of fit. The following equations were used to calculate the RMSE and MAE:

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^N |R_i - P_i|} \quad (1)$$

$$MAE = \frac{1}{N} |R_i - P_i| \quad (2)$$

Where R_i and P_i denote as the ground truth and predicted HR and BR, respectively. While N is the total number of heartbeats and respiration rate per minute.

IV. RESULT AND DISCUSSION

A. Datasets

This work used a self-collected dataset of 80 videos, as described in [20]. A total of 20 subjects with frontal view of face videos and 4 videos from each participant were obtained. All subjects are in good health and have signed a consent form to participate in the study. The video was shot under visible lighting with a webcam and a smartphone camera. Subjects perform three different conditions: 1) relaxed mode, (2) after a walking exercise, and (iii) after going up and down stairs. After each exercise, subjects were asked to sit still, and their facial expressions were recorded. The ground truth of each subject's HR and BR were taken within 60 seconds, with the pulse sensor and pulse oximeter collecting the data and

feeding it to the Arduino microcontroller for data evaluation. The final heartbeat values were displayed on the LCD screen.

Fig. 4 shows the schematic design of the prototype to collect the benchmark value of the HR and BR. The development and testing of this prototype has been done using standard PM100 pulse oximeter model. This prototype of self-collected facial dataset will be tested with another state-of-the-art algorithm to extract the rPPG signal using CNN-based model. The CNN network was implemented using Phyton with Tensor Flow and Keras framework within Colab notebook. The proposed work is validated with our self-collection post-exercise dataset which contains facial videos of different lighting conditions and taken from smartphone or webcam. We have produced three different set of exercise mode to evaluate the proposed framework efficiency. At the same time, validate the effectiveness of our health monitoring prototype. We observed that the dataset which consists varying pose of less motion yields better accuracy in terms of HR and BR estimation. Motion artefacts and lighting variations also causing lots of noise in the extracted rPPG signal. Hence, the number of false positive or negative peak detections will increase.

B. Heart Rate Analysis

The simulation work was done in order to compare the measured HR to the predicted HR. The comparisons for the rest and post-exercise conditions are shown in the following figures. The efficacy of the proposed RoI was experimentally validated using the extracted rPPG signal for both HR and BR estimation. Fig. 5 depicts a comparison of measured and estimated HR when the subject is at rest. The estimated HR falls within the normal range of 60 to 100bpm. Fig. 6 and 7 depict a sample of the subject after exercises, which are walking and staircase exercises. The difference between measured and estimated HR for post-exercise falls within a range of more than 80 bpm. According to our observations, the estimated HR based on the CNN model agrees well with the measured HR from the prototype developed in our previous work. Our proposed method accurately calculates the changes in estimated HR from rest to post exercise.

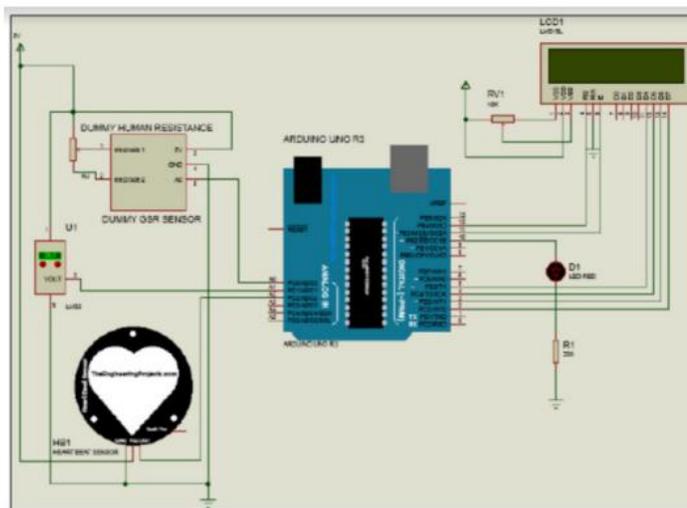


Fig. 4. The Schematic Design of the Health Monitoring Prototype [20].

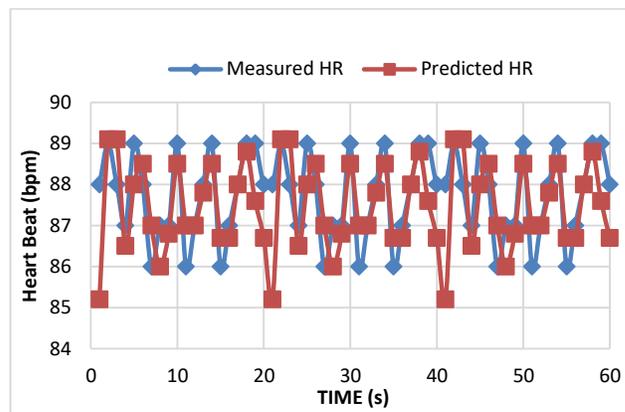


Fig. 5. Measured Heart Rate against the Estimated Heart Rate while Subjects Resting Conditions.

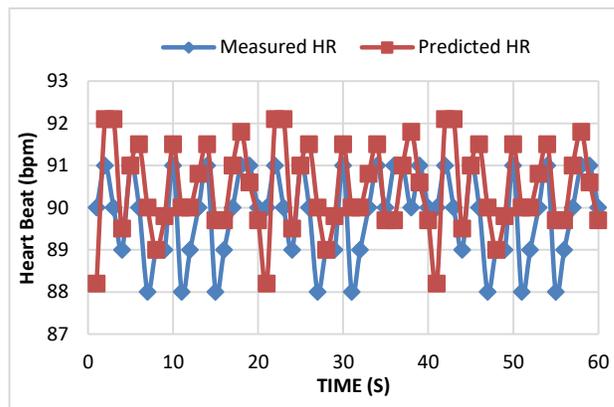


Fig. 6. Measured Heart Rate against the Estimated Heart Rate for Subject after Walking Exercise.

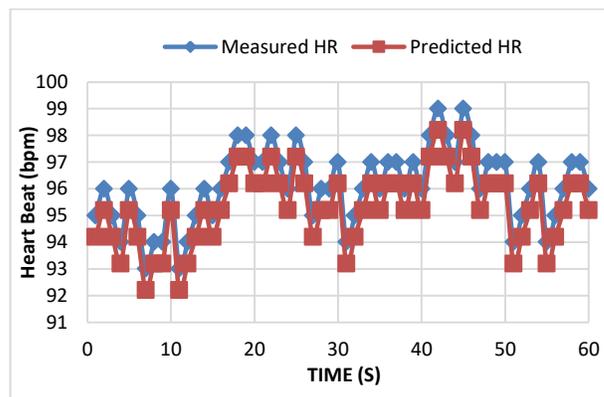


Fig. 7. Measured Heart Rate against the Estimated Heart Rate for Subject after Stairs Exercise.

From the results in Table III, the RMSE and MAE are derived from the average measured of the actual HR and predicted HR for all subjects for several types of exercises. The average of actual HR for resting, walking and after stairs exercise are 87.75, 89.8 and 95.3, respectively. The MAE for our proposed method using our self-collected physical exercise dataset is very good, falling within the 2.16 ± 2.2 beats/min range. To evaluate our dataset using our proposed method, we compared it to other state-of-the-art methods based on CHROM, ICA, PCA and HR-CNN. Our proposed

method outperforms the traditional CHROM, ICA and PCA methods with low MAE and RMSE. We conclude that CNN based methods are more accurate and present relevant beats from the pulse rate series for HR estimation. We demonstrate the similar process for BR performance.

TABLE III. HR ANALYSIS BASED ON AVERAGED MEASURE, RMSE AND MAE

Method	Exercise	Predicted _{avg} HR	RMSE	MAE
CHROM[7]	Rest	75.97	11.47	5.4
	Walk	78.98	10.92	7.64
	Stairs	69.6	21.19	9.48
ICA[15]	Rest	75.2	5.22	0.97
	Walk	67.65	12.67	8.21
	Stairs	68.8	14.94	10.05
PCA[15]	Rest	85.43	9.72	3.9
	Walk	77.88	2.17	1.14
	Stairs	79.03	7.44	7.98
HR-CNN[11]	Rest	83.05	2.52	2.80
	Walk	80.65	2.55	2.91
	Stairs	88.37	2.37	2.15
Proposed Method	Rest	87.47	1.92	1.72
	Walk	80.48	2.37	2.96
	Stairs	86.1	2.64	1.8
	Overall		2.31	2.16

C. Breathing Rate Analysis

Fig. 8 depicts a comparison of measured and estimated BR when the subject is at rest. The estimated BR falls within the normal range of 15 to 20bpm. Fig. 9 and 10 show a sample of the subject after exercises, which are walking and stairwell exercises. The difference between measured and estimated BR for post-exercise falls within a range of more than 20 bpm. The curves are simulated from the extracted rPPG signals within a constant distance from the camera. ow SNR. Rapid changes of breathing rate can be detected from resting position and stairs exercise. This is accomplished by selecting the peak in the spectrum that provides the highest SNR for the pulse signal.

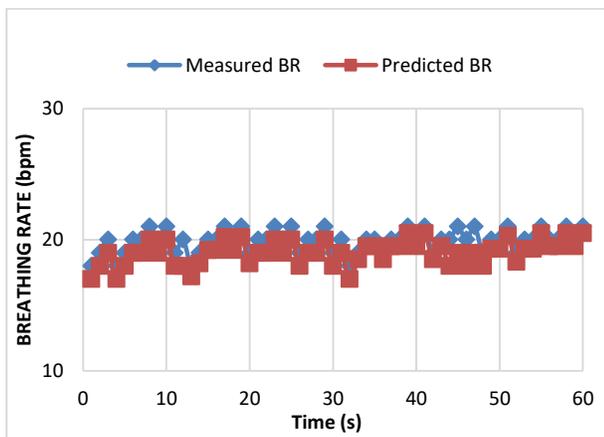


Fig. 8. Measured Heart Rate against the Estimated Heart Rate while Subjects at Resting Conditions.

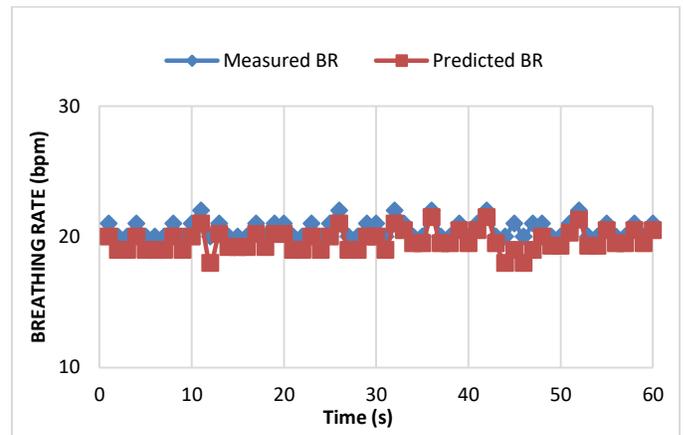


Fig. 9. Measured Heart Rate against the Estimated Heart Rate for Subject after Walking Exercise.

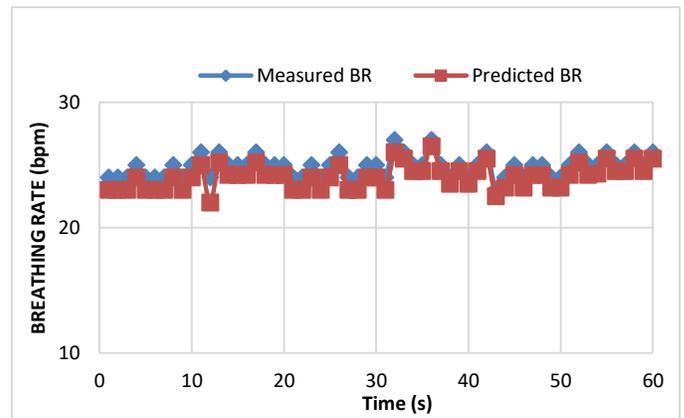


Fig. 10. Measured Heart Rate against the Estimated Heart Rate for Subject after Stairs Exercise.

Table IV derived the RMSE and MAE from the average measured of the actual BR and predicted BR for all subjects for several types of exercises. The average of actual BR for resting, walking and after stairs exercise are 19.9, 20.5 and 24.8, respectively. The MAE for our proposed method of estimating BR is between 2.16 and 2.2 beats/min. We demonstrated the effectiveness of BR peak finding interpolation and reported an acceptable MAE of less than 2 bpm. The method is compared to various state-of-the-art methods (CHROM, PCA, and ICA) and an HR-CNN-based method. For our specific post-exercise dataset, our proposed technique produced valuable findings. Although the use of the CNN framework in our suggested method does not compete with other deep learning methods, it is sufficient to examine the practicality of our prototype as a non-invasive health assessment system.

The results presented for HR and BR estimation show a high level of agreement between calculated and ground truth measurements. Based on the HR and BR analysis in Table III and Table IV, the CHROM method has the highest RMSE and MAE. CHROM is typically used to remove noise caused by light reflection via colour difference channel normalisation. For the purposes of comparison, both the ICA and PCA methods are types of BSS techniques in which independent signal sources are filtered from a mixed signal. Both ICA and

PCA improved the RMSE and MAE in this simulation work. Most rPPG research studies have a very limited post-exercise dataset. Despite the fact that several deep learning methods have been implemented in the facial-based rPPG dataset, our results validated on this self-collected post-exercise dataset are comparable to other state-of-the-art algorithm.

Based on the experimental results and performance evaluation, the proposed RoI of facial regions is effective for CNN model in forming spatio-temporal features for rPPGg signal estimation. A low percentage of MAE and RMSE indicates that the RoI selection is significant in improving the SNR for better signal quality assessment. We can assume that the majority of significant pulse pixels are located on forehead and cheeks. In practice, the most significant impediment to accurate HR and BR analysis is false peak detection. The signal was interpolated at 256 Hz to sharpen the peaks and manage the latency. Beat-to-beat pulse rate values were computed from the interbeat intervals. When calculating the interval between beats, a false peak detection appears as an incorrect beat and causes a major error in the HR and BR analysis of a healthy person. Removing unnecessary peaks may also have an impact on HR and BR estimation.

However, there are few limitations included in our study. The HR and BR estimation from the extracted rPPG signal requires further evaluation using CNN. The end-to-end approach of CNN model will be efficiently incorporated into our model's future development. Due to the limitations of the dataset used for this study, we will increase the duration of exercise and the recorded time to produce more variation in the HR and BR signal patterns. In the future, we will use adaptive RoI detection to improve the efficacy of our region-based skin detection model.

TABLE IV. BR ANALYSIS BASED ON AVERAGED MEASURE, RMSE AND MAE

Method	Exercise	Predicted _{avg} BR	RMSE	MAE
CHROM[7]	Rest	17.28	14.98	4.46
	Walk	17.85	7.99	3.46
	Stairs	22.30	13.77	2.38
ICA[15]	Rest	18.28	2.16	1.01
	Walk	18.85	3.17	1.01
	Stairs	23.3	2.95	3.93
PCA[15]	Rest	18.73	4.08	2.97
	Walk	19.30	5.09	2.97
	Stairs	23.75	2.87	1.89
HR-CNN[11]	Rest	18.52	2.53	1.89
	Walk	20.25	1.94	1.85
	Stairs	25.61	1.52	1.51
Proposed Method	Rest	19.03	2.93	0.89
	Walk	19.60	0.94	1.89
	Stairs	24.05	1.72	1.81
	Overall		1.86	1.53

VI. CONCLUSION AND FUTURE WORK

We may conclude that the results are promising enough to support the effectiveness of the CNN model in extracting relevant pulse pixels for further analysis. For HR estimation, our methods achieved an average RMSE of 2.31 and MAE of 2.16. The overall BR estimation had an average RMSE of 1.86 and an MAE of 1.53. Our methods perform well in a variety of post-exercise facial video streams under controlled lighting conditions, whether utilizing a camera or a smartphone.

In the future, an end-to-end CNN approach will be established using this dataset for simulating HR and BR estimation. This system can be upgraded to detect sudden changes in heart rate and breathing patterns. Furthermore, this system made use of a partial face region, which was expected to contribute the most blood variation. The system will be improved in real-time scenario especially when subject performing head movements.

Other factors under consideration for future research include increase the number of hidden layers in the CNN framework and optimizing the network design to achieve more promising outcomes. The number of layers, number of filters per convolutional layer, and number of neurons per dense layer could all have a significant impact and provide an automatic method of determining the best network architecture. Further investigation can be performed such as improving the CNN framework and comparing the influence of colour channel performance, particularly in terms of rPPG signal accuracy and artifact removal.

ACKNOWLEDGMENT

This research was supported by Universiti Tun Hussein Onn Malaysia (UTHM) through Tier 1 Grant (Vot H756).

REFERENCES

- [1] Jean Pierre Tincopal , Paulo Vela-Anton, Cender U. Quispe-Juli, Anthony Arostegui, "Development of an IoT Device for Measurement of Respiratory Rate in COVID-19 Patients", In International Journal of Advanced Computer Science and Applications, pp. 1009-1016, Vol. 13, No. 4, 2022.
- [2] Chun-Hong Cheng, Kwan-Long Wong, Jing-Wei Chin et al., "Deep Learning Methods for Remote Heart Rate Measurement: A Review and Future Research Agenda", In Sensors (Basel, Switzerland), Vol. 21, No. 18, 2020.
- [3] Ming-Zher Poh, Daniel J. McDuff, Rosalind W. Picard, "Non-contact, automated cardiac pulse measurements using video imaging and blind source separation", In Optics Express, pp. 10762-10774, Vol. 18, No. 10, 2010.
- [4] Poh, M.Z., McDuff, D.J., and Picard, R.W, "Advancements in noncontact, multiparameter physiological measurements using a webcam", In IEEE Transactions on Biomedical Engineering, pp. 7-11, Vol. 58, No. 1, 2010.
- [5] Juan Cheng, Xun Chen, Lingxi Xu. et al., "Illumination Variation-Resistant Video-Based Heart Rate Measurement Using Joint Blind Source Separation and Ensemble Empirical Mode Decomposition," In IEEE Journal of Biomedical and Health Informatics, pp. 1422-1433, Vol. 21, No. 5, 2017.
- [6] Wenjin Wang, Albertus C. den Brinker, Sander Stuijk et al., "Algorithmic Principles of Remote PPG," In IEEE Transactions on Biomedical Engineering, pp. 1479-1491, Vol. 64, No. 7, 2017.
- [7] Gerard de Haan, Vincent Jeanne, "Robust Pulse Rate from Chrominance-Based rPPG," In IEEE Transactions on Biomedical Engineering, pp. 2878-2886, Vol. 60, No. 10, 2013.

- [8] Yu-Chen Lin, Nai-Kuan Chou, Guan-You Lin et al., "A Real-Time Contactless Pulse Rate and Motion Status Monitoring System Based on Complexion Tracking," In *Sensors (Switzerland)*, Vol. 17, No. 7, 2017.
- [9] Ming-Zher Poh, Daniel J. McDuff, Rosalind W. Picard, "Advancements in Noncontact, Multiparameter Physiological Measurements Using a Webcam," In *IEEE Transactions on Biomedical Engineering*, pp. 7-11, Vol. 58, No. 1, 2011.
- [10] Giovanni Cennini, Jeremie Arguel, Kaan Aksit et al., "Heart Rate Monitoring Via Remote Photoplethysmography with Motion Artifacts Reduction," In *Optics Express*, pp. 4867-4875, Vol. 18, No. 5, 2010.
- [11] Radim Spetlik, Jan Cech, Vojtěch Franc et al., "Visual Heart Rate Estimation with Convolutional Neural Network," In the *British Machine Vision Conference*, 2018.
- [12] Ying Qiu, Yang Liu, Juan Arteaga-Falconi et al., "EVM-CNN: Real-Time Contactless Heart Rate Estimation from Facial Video," In *IEEE Transactions on Multimedia*, pp. 1778-1787, Vol. 21, No. 7, 2019.
- [13] Weixuan Chen, Daniel McDuff, "Deepphys: Video-Based Physiological Measurement Using Convolutional Attention Networks," In *Proceedings of the European Conference on Computer Vision*, pp. 349-365, 2018.
- [14] Rencheng Song, Senle Zhang, Chang Li et al., "Heart Rate Estimation from Facial Videos Using a Spatiotemporal Representation with Convolutional Neural Networks," In *IEEE Transactions on Instrumentation and Measurement*, pp. 7411-7421, Vol. 69, No. 10, 2020.
- [15] Chuanxiang Tang, Jiwu Lu, Jie Liu, "Non-contact Heart Rate Monitoring by Combining Convolutional Neural Network Skin Detection and Remote Photoplethysmography via a Low-Cost Camera," In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 1309-1315, 2018.
- [16] Mohammad Sabokrou, Masoud Pourreza, Xiaobai Li, Mahmood Fathy, Guoying Zhao, "Deep-HR: Fast Heart Rate Estimation from Face Video Under Realistic Conditions," In *Expert System with Applications*, 2020.
- [17] Xingjian Shi, Zhoung Chen, Hao Wang et al, "Convolutional LSTM Network: A Machine Learning Approach for Precipitation Nowcasting," In *Proc. of 28th International Conference on Neural Information Processing Systems*, pp. 802-810, Vol. 1, 2015.
- [18] Yun-Yun Tsou, Yi-An Lee, Chiou-Ting Hsu et al, "Siamese-rPPG network: remote photoplethysmography signal estimation from face videos," In *proceedings of the 35th Annual ACM Symposium on Applied Computing*, pp. 2066-2073, 2020.
- [19] Zhan Q, Wang W, de Haan G, "Analysis of CNN-based remote-PPG to understand limitations and sensitivities," In *Biomedical Optics Express*, pp. 1268-1283, Vol. 11, No. 3, 2020.
- [20] Nor Surayahani Suriani, Nur Adlina Jumain, Abdalla Abdurahman Ali, "Facial Video based Heart Rate Estimation for Physical Exercise," In *IEEE Symposium on Industrial Electronics & Applications*, pp.1-5, 2021.
- [21] Azizullah Kakar, Naveed Sheikh, Bilal Ahmed, Saleem Iqbal, Abdul Rahman, Saboor Ahmad Kakar, Arbab Raza, Samina Naz, Junaid Baba, "Systematic Analysis and Classification of Cardiac Rate Variability using Artificial Neural Network," In *International Journal of Advanced Computer Science and Applications*, pp 746-750, Vol.9, No.11, 2018.
- [22] Dae-Yeol Kim, Kwangkee Lee, Chae-Bong Sohn, "Assessment of ROI Selection for Facial Video-Based rPPG," In *Sensors (Basel)*, ISSN: 1424-8220, pp. 7923, Vol. 21, No. 23, 2021.
- [23] D. Datcu, M. Cidota, S. Lukosch and L. Rothkrantz, "Noncontact automatic heart rate analysis in visible spectrum by specific face regions", In *ACM International Conference Proceeding Series*, Vol. 767, 2013.
- [24] S. Kwon, J. Kim, D. Lee and K. Park, "ROI analysis for remote photoplethysmography on facial video", In *Proc. 37th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, pp. 4938-4941, 2015.
- [25] J. Chen and A. Flabouris, "Respiratory Rate: The Neglected Vital Sign," In *The Medical Journal of Australia*, Vol.188, No. 11, pp. 657, 2008.

A Survey on MCT vs. DCT: Who is the Winner in COVID-19

Omar Khattab

Department of Computer Science and Engineering
Kuwait College of Science and Technology (KCST), Doha, Kuwait

Abstract—Coronavirus disease (COVID-19) is a contagious disease appeared in late 2019 and caused by a virus called SARS-CoV-2. It is a pandemic spreading across the whole world and impacts millions of people and sadly causes death. There are two main Contact Tracing Methods (CTMs) to limit and slow down any chance of transmission of it: Manual Contact Tracing (MCT) and Digital Contact Tracing (DCT). The MCT abides by the guide to World Health Organization's guidance (WHO) on COVID-19 in terms of properly applying social distancing, wearing masks, washing hands, using sanitizers, etc. while the DCT abides by the digital contact tracing applications developed by several countries. This survey is mainly focused on these CTMs and the recent proposed solutions in this field, in order to highlight their drawbacks that negatively impact on both of satisfaction and feasibility in using them. The findings in the survey will be beneficial to understand the effectiveness of CTMs and current proposed solutions, in order to develop a comprehensive smart tracking system able to cooperatively contribute with both of MCT and DCT in extremely detecting, preventing, and slowing down the spread of COVID-19 or even any other similar pandemics in the future.

Keywords—COVID-19; coronavirus disease; manual contact tracing; digital contact tracing

I. INTRODUCTION

Coronavirus disease (COVID-19) has been emerged for the first time in Wuhan [1-9], where the first reported cases were found in Huanan seafood market [10-12]. It has been reported by the World Health Organization (WHO) that there are 318,648,834 confirmed cases of COVID-19, including 5,518,343 deaths [13]. To slowing the spread of COVID-19 and protect family and community, Contact Tracing Methods (CTMs): Manual Contact Tracing (MCT) and Digital Contact Tracing (DCT) play a vital role in this respect, The MCT is a manual method [14-20] strives to [21]:

- 1) Support COVID-19 patients to stay home and self-isolate.
- 2) Alert and help people who have been in close contact with COVID-19 patients.
- 3) Follow-up them in testing, quarantine, and wearing a mask properly.

The DCT is a dynamic method [22-34] based on smartphones' applications which strives in tracking people diagnosed with COVID-19 to notify mobile users whether they

have been exposed to the virus or not. The DCT utilize several technologies to collect data [35]: cell tower location data, Quick Response (QR) code, credit card/public transit card, videos surveillance, Global Positioning System (GPS), and Bluetooth.

There are three main approaches to store users' sensitive data [36]: Centralized Approach (CA) where the data is stored on centralized servers, Decentralized Approach (DA) where the data is stored on individual mobiles, and Hybrid Approach (HA) which is combination between CA and DA. The DA aims to protect users' sensitive data where a central server plays a small role in this respect by keeping very little data [37]. The CA is more efficient [37] and more secure compared with the DA, while it is considered a single point of failure [38].

To the best of our knowledge, existing related works have not fully considered CTMs and recent proposed solutions. Therefore, this paper thoroughly surveys both of CTMs and recent proposed solutions to highlight their drawbacks that negatively impact on both of satisfaction and feasibility in using them. The rest of the paper is organized as follows: In Section II, an overview of CTMs and proposed solutions is presented: DCT and MCT. In Section III, a comprehensive comparison of CTMs and proposed solutions is presented: MCT vs. DCT, DCT approaches, and DCT proposed solutions. In Section IV, a discussion is presented. Finally, a conclusion is given in Section V.

II. OVERVIEW OF CTMS AND PROPOSED SOLUTIONS

To provide a comprehensive comparison of CTMs, this section presents lots of vital recent related works either theoretical or practical: surveys, overviews, and proposed solutions for DCT and MCT.

A. DCT

In [22], the authors have introduced a Self-Sovereign Identity (SSI) model based blockchain to address the following issues in DCT applications: privacy leakage, efficiency, and energy consumption. The effectiveness of the proposed solution has been validated by theoretical analysis.

In [27], the authors have proposed a blockchain enabled privacy preserving contact tracing scheme: BeepTrace. Numerical analysis has shown higher security and privacy, battery friendly, and globally accessible.

In [34], a framework of a blockchain, Artificial Intelligence (AI) and Internet of Things (IoT) based system for the detection of COVID-19 and distancing has been proposed. It has aimed to offer real time data sharing, security, and transparency. However, no implementation or testing provided about the framework.

In [35], an overview of several DCT applications has been conducted. It has been concluded that the DCT have had the following issues: limited Internet access in poor countries, people without smartphones, lack of signal, transparency, privacy and security.

In [39], a cross national online survey has been conducted in the UK, Republic of Ireland, and the US, in order to discover public attitudes and the acceptability of DCT. It has been concluded that that trust and privacy have been the main concerns for the adoption of DCT.

In [40], 41 countries and 23 US states have developed a total of 64 DCT applications, where they have sampled eight applications in European countries between British Isles, and mainland Europe equitably. Using various analysis (e.g., quantitative analysis and qualitative coding), it has been concluded that the DCT has required updating regularly due to governmental policies and guidelines, there have been issues in the devices, applications and their features, high battery consumption, usability should be enhanced, and users have been generally unhappy with the applications.

In [41], a detailed analysis of DCT applications for 32 countries has been presented. The proposed architecture using blockchain Hyperledger Fabric (HF) has addressed the following inherent issues related to contact tracing: security, privacy, authentication, access control, flexibility, scalability, interoperability, and efficiency.

In [42], the authors have proposed COVERT blockchain HF for COVID-19 contact tracing with keeping user's privacy. The results have proved its scalability, robustness, and efficiency in protecting privacy leakage.

In [43], the authors have proposed a blockchain platform for contact tracing with keeping user's privacy, using a Generative Adversarial Network (GAN) application. It has shown that the privacy has been addressed by iterative deleting older data from the database.

In [44], the authors have proposed a prototype of blockchain and SSI-based digital contact tracing platform, using Mystiko blockchain cluster. A performance evaluation of the platform has been conducted, where it has shown addressing issues in security, privacy, scalability, and transaction throughput features.

In [45], the authors have proposed a framework used off-chain scaling mechanism of Interplanetary File System (IPFS) for contact tracing. There a performance evaluation using Ethereum application has been conducted of the framework in terms of security, privacy, and scalability.

In [46], the authors have proposed and implemented a blockchain based system called COVID-19 Contact Tracing System (CCTS), to verify, track and detect new cases of

COVID-19, using Ethereum application. However, user's privacy and accuracy in detecting contacts of COVID-19 have not been investigated.

The DCT provides several features: accurate, fast, and low cost [47].

B. MCT

It is a traditional contact tracing method manually managed by health care providers to identify the contacts of infected individuals, interview, alert them to quarantine, and to seek a test [47]. However, the MCT has become difficult to be used due to rapid spread of COVID-19 [48], as it has been some drawbacks: relying on human memory, taking time [49], requiring trained human resources [47], [50], inefficient [51], costly, highly error prone, and not scalable [14].

The recent studies in [52] and [53] shows that the combination of MCT and DCT is more efficient for contact tracing.

III. COMPARISON OF CTMS AND PROPOSED SOLUTIONS

In Section II, plenty of recent related works have been considered: surveys, overviews, and proposed solutions of CTMs. To provide a comprehensive comparison of CTMs, in this Section, three types of comparisons are presented: MCT vs. DCT, DCT approaches, and DCT proposed solutions.

A. MCT vs. DCT

Ten factors are considered to distinguish between CTMs (MCT and DCT): time, efficiency, accuracy, cost, diagnosis, failure, scalability, reliability, dependency, and investigation. This is shown in Table I. In this competition, the DCT obviously has a full advantage over the MCT.

B. DCT Approaches

Five factors are considered to distinguish between the DCT approaches (CA, DA, and HA): data storage location, efficiency, privacy, security, and point of failure. This is shown in Table II. In this competition, the HA obviously has dominant features over the CA and DA.

TABLE I. MCT vs. DCT

Comparison	DCT	MCT
Time	Less	More
Efficiency	More	Less
Accuracy	More	Less
Cost	Less	More
Diagnosis	Fast	Slow
Failure	Low	High
Scalability	High	Low
Reliability	High	Low
Dependency	Technology	Human memory
Investigation	GPS, QR code, Bluetooth cell tower location data, credit card/public transit card, videos surveillance	Self-assessment survey

TABLE II. DCT APPROACHES

Comparison	CA	DA	HA
Data storage location	Server (S)	Mobile (M)	S & M
Efficiency	High	Low	Average
Privacy	Low	High	Average
Security	High	Low	Average
Point of failure	High	Low	Average

TABLE III. DCT PROPOSED SOLUTIONS

Paper	Year	Type of Research	Implementation	Concern	Addressing	Solution
[22]	2021	Theoretical	Analysis	n/a	Privacy, efficiency, energy consumption	SSI model based blockchain
[27]	2021	Theoretical	Analysis	n/a	Security, privacy, battery friendly, globally accessible	BeepTrace based blockchain
[34]	2021	Theoretical	n/a	Not developed	Real time data sharing, security, transparency	Framework based blockchain, AI and IoT
[35]	2020	Theoretical	Overview	Limited Internet access, people without smartphones, lack of signal, transparency, privacy, security	n/a	n/a
[39]	2021	Theoretical	Survey	Trust and privacy	n/a	n/a
[40]	2021	Theoretical	Analysis	Updating DCT regularly, issues in the devices, applications and their features, battery consumption, usability, user satisfaction	n/a	n/a
[41]	2021	Practical	HF	-	Security, privacy, authentication, access control, flexibility, scalability, interoperability, efficiency	Architecture based blockchain
[42]	2021	Practical	HF	-	Scalability, robustness, privacy	COVERT based blockchain
[43]	2021	Practical	GAN	-	Privacy	Platform based blockchain
[44]	2021	Practical	Mystiko	-	Security, privacy, scalability, transaction throughput features	SSI model based blockchain
[45]	2021	Practical	Ethereum	-	Security, privacy, scalability	Framework based blockchain and IPFS
[46]	2021	Practical	Ethereum	Privacy, accuracy	Verify, track, detect new cases of COVID-19	CCTS based blockchain

C. DCT Proposed Solutions

Five factors are considered to distinguish between the DCT proposed solutions [22, 27, 34, 35, 39-46]: type of research, implementation, concern, addressing, and solution. This is shown in Table III.

As for the type of research, it can be seen that the related works have been equitably conducted between theoretical and practical research works. This is shown in Fig. 1.

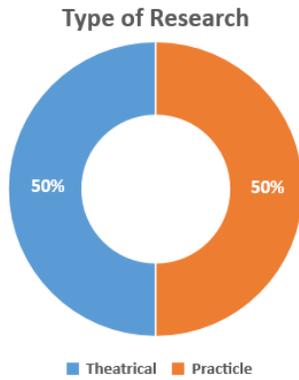


Fig. 1. Type of Research for DCT Proposed Solutions.

For the implementation, three related works are used analysis, followed by HF and Ethereum with two research works each; lastly, overview, survey, GAN, and Mystiko. This is shown in Fig. 2.

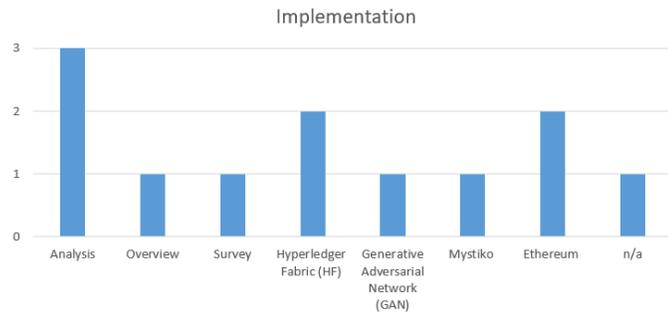


Fig. 2. Implementation for DCT Proposed Solutions.

In terms of the concern, fifteen common issues are arisen and divided into two main categories: IT's issues and user's issues. This is shown in Fig. 3.

In the addressing, eighteen common issues are arisen. It has been noticed that all the addressed issues are related to IT issues. This is shown in Fig. 4.

Finally, the most DCT proposed solutions based on the blockchain, as shown in Fig. 5.

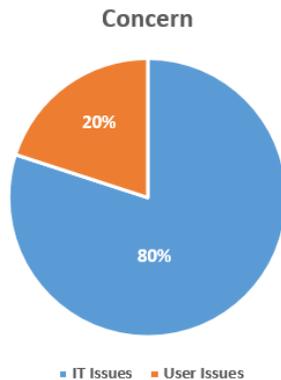


Fig. 3. Concern for DCT Proposed Solutions.

Addressing

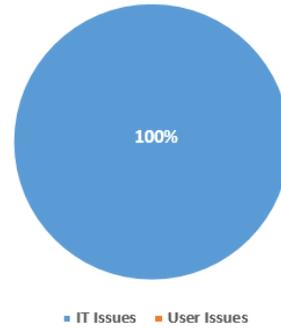


Fig. 4. Addressing for DCT Proposed Solutions.

Solution

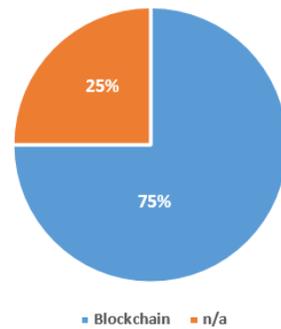


Fig. 5. DCT Proposed Solutions

IV. DISCUSSION

In Section III, a comparison of CTMs has been conducted, where three types of comparisons have been presented: MCT vs. DCT, DCT approaches, and DCT proposed solutions.

For MCT vs. DCT, the DCT has shown a full advantage over the MCT in terms of time, efficiency, accuracy, cost, diagnosis, failure, scalability, reliability, dependency, and investigation, as shown in Table I.

For DCT approaches, the HA has shown dominant features over the CA and DA in terms of data storage location, efficiency, privacy, security, and point of failure, as shown in Table II.

For DCT proposed solutions, twelve research works have been conducted [22, 27, 34, 35, 39-46]. It has been noticed that all these works have been confined in introducing, enhancing or proposing DCT applications related to IT issues and user issues, as shown in Table III.

Therefore, in addition to the combination between DCT (HA approach) and MCT, it would be more effective to propose, implement and, distribute a comprehensive smart tracking system located in public places, universities, schools, hospitals, banks, airports etc. This obviously will extremely limit and slow down any chance of transmission of COVID-19 or even any other similar pandemics in the future. This is shown in Fig. 6.

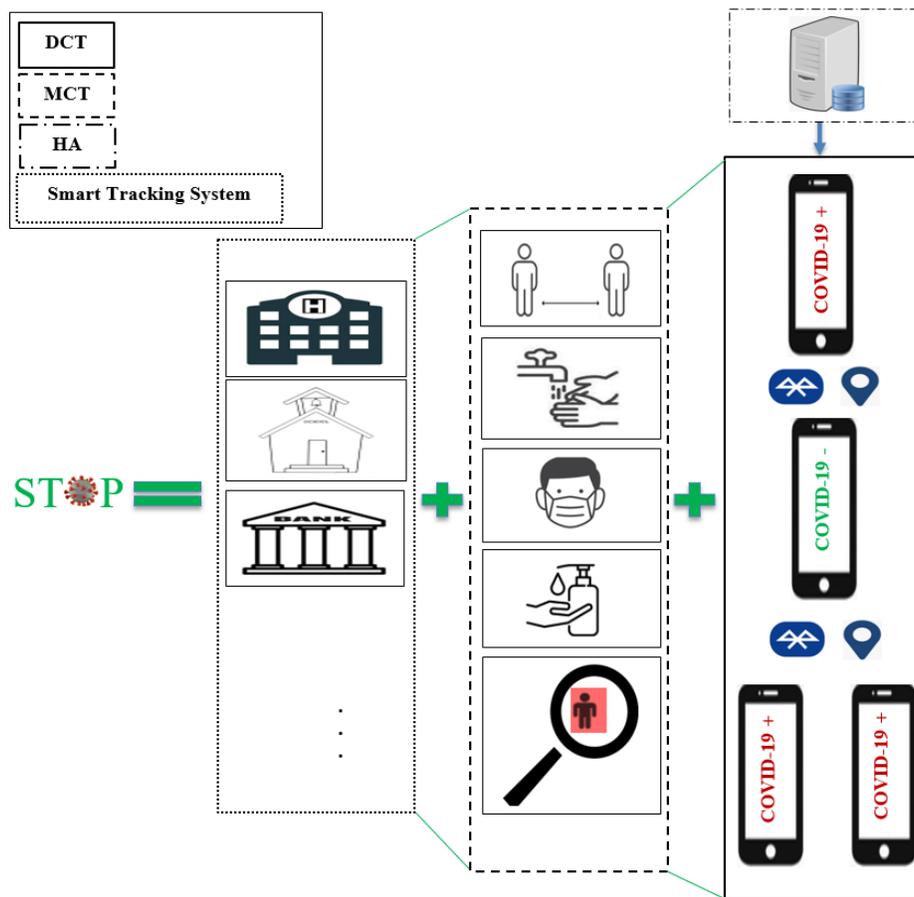


Fig. 6. Comprehensive CTM.

V. CONCLUSION

In this paper, the CTMs and the recent proposed solutions for COVID-19 have been surveyed thoroughly, where a fair comparison has been presented: MCT vs. DCT, DCT approaches, and DCT proposed solutions. For DCT proposed solutions, twelve research works have been conducted. It has been concluded that all these works have been confined in introducing, enhancing or proposing DCT applications related to IT issues and user issues.

Therefore, the competition in this paper has shown the importance of using both of MCT and DCT (HA approach), and come up with a comprehensive smart tracking system able to cooperatively contribute with them in extremely detecting, preventing, and slowing down the spread of COVID-19 or any other similar pandemics in the future.

REFERENCES

- [1] M. Fradi and M.Machhout, "Real-time application for Covid-19 class detection based CNN architecture," IEEE International Conference on Design & Test of Integrated Micro & Nano-Systems (DTS), Sfax, Tunisia, pp. 1-6, 2021.
- [2] Y. Jiang, H. Chen, M. Loew and H. Ko, "Covid-19 CT image synthesis with a conditional generative adversarial network," IEEE Journal of Biomedical and Health Informatics, vol. 25, no. 2, pp. 441-452, 2021.
- [3] W. Haochen, "Western perspectives of the Covid-19 in China," International Conference on Public Health and Data Science (ICPHDS), Guangzhou, China, pp. 116-119, 2020.
- [4] M. Wang, W. Kong, J. Xie and S. Xu, "Modeling the Covid-19 epidemic in PR China," 7th International Conference on Big Data and Information Analytics (BigDIA), Chongqing, China, pp. 324-333, 2021.
- [5] S. Vishnu and S. Jino Ramson, "An internet of things paradigm: pandemic management (incl. Covid-19)," International Conference on Artificial Intelligence and Smart Systems (ICAIS), Coimbatore, India, pp. 1371-1375, 2021.
- [6] P. Porwal, K. Thirunavukkarasu, A. Sinha and A. Singh, "Data analysis and detection of coronavirus disease using convolution neural network," 2nd International Conference on Advances in Computing, Communication Control and Networking (ICACCCN), Greater Noida, India, pp. 786-790, 2020.
- [7] H. Turabieh and W. Ben Abdesslem, "Predicting the existence of Covid-19 using machine learning based on laboratory findings," International Conference of Women in Data Science at Taif University (WiDSTaif), Taif, Saudi Arabia, pp. 1-7, 2021.
- [8] L. Zhu, W. Dong, Q. Sun, E. Vargas and X. Du, "Estimation of the unreported infections of Covid-19 based on an extended stochastic susceptible-exposed-infective-recovered model," IEEE 10th Data Driven Control and Learning Systems Conference (DDCLS), Suzhou, China , pp. 953-958, 2021.
- [9] T. Karakose, TY. Ozdemir, S. Papadakis, R. Yirci, SE. Ozkayran and H. Polat, "Investigating the Relationships between COVID-19 Quality of Life, Loneliness, Happiness, and Internet Addiction among K-12 Teachers and School Administrators—A Structural Equation Modeling Approach," International Journal of Environmental Research and Public Health, vol. 19, no. 3, pp. 1-20, 2022.
- [10] W. Yang, Q. Cao, L. Qin, X. Wang, Z. Cheng, A. Pan, J. Dai, Q. Sun, F. Zhao, J. Qu and F. Yan, "Clinical characteristics and imaging manifestations of the 2019 novel coronavirus disease (Covid-19): A

- multi-center study in Wenzhou city, Zhejiang, China,” *J Infect*, vol. 80, no. 4, pp. 388-393, 2020.
- [11] L. Gralinski and V. Menachery, “Return of the coronavirus: 2019-nCoV,” *Viruses*, vol. 12, no. 2, pp. 1-8, 2020.
- [12] D. Kuldeep, K. Sharun, T. Ruchi, S. Shubhankar, B. Sudipta, M. Yashpal, S. Karam, C. Wanpen, A. Katterine and M. Alfonso, “Coronavirus disease 2019– Covid-19,” *Clinical Microbiology Reviews*, vol. 33, no. 4, pp. 1-48, 2020.
- [13] World Health Organization. (Jan, 2022). WHO coronavirus (Covid-19) Dashboard. Available: <https://covid19.who.int/>.
- [14] M. Chowdhury, M. Ferdous, K. Biswas, N. Chowdhury and V. Muthukkumarasamy, “Covid-19 contact tracing: challenges and future directions,” *IEEE Access*, vol. 8, pp. 225703-225729, 2020.
- [15] P. Ng, P. Spachos, S. Gregori and K. Plataniotis, “Personal devices for contact tracing: smartphones and wearables to fight Covid-19,” *IEEE Communications Magazine*, vol. 59, no. 9, pp. 24-29, 2021.
- [16] M. Bano, C. Arora, D. Zowghi and A. Ferrari, “The rise and fall of Covid-19 contact-tracing apps: when NFRs collide with pandemic,” *IEEE 29th International Requirements Engineering Conference (RE)*, Notre Dame, IN, USA, pp. 106-116, 2021.
- [17] P. Ng, P. Spachos and K. Plataniotis, “Covid-19 and your smartphone: BLE-based smart contact tracing,” *IEEE Systems Journal*, vol. 15, no. 4, pp. 5367-5378, 2021.
- [18] B. Patel, N. Jain, R. Menon and S. Kodeboyina, “Comparative study of privacy preserving-contact tracing on digital platforms,” *International Conference on Computational Intelligence (ICCI)*, Bandar Seri Iskandar, Malaysia, pp. 137-141, 2020.
- [19] A. Khandelwal, A. Kotwal, P. Sutone and V. Wag, “Automated contact tracing using person tracking and re-identification,” *2nd International Conference on Secure Cyber Computing and Communications (ICSCCC)*, Jalandhar, India, pp. 102-107, 2021.
- [20] D. Yazti and C. Claramunt, “Covid-19 mobile contact tracing apps (MCTA): a digital vaccine or a privacy demolition?,” *21st IEEE International Conference on Mobile Data Management (MDM)*, Versailles, France, pp. 1-4, 2020.
- [21] Centers for Disease Control and Prevention. (Feb, 2022). Contact tracing. Available: <https://www.cdc.gov/coronavirus/2019-ncov/daily-life-coping/contact-tracing.html#>.
- [22] D. Wang, X. Chen, L. Zhang, Y. Fang and C. Huang, “A blockchain based human-to-infrastructure contact tracing approach for Covid-19,” *IEEE Internet of Things Journal (Early Access)*, pp.1-1, 2021.
- [23] H. Faria, S. Paiva and P. Pinto, “An advertising overflow attack against android exposure notification system impacting Covid-19 contact tracing applications,” *IEEE Access*, vol. 9, pp. 103365-103375, 2021.
- [24] G. BetarteJ. Campo, A. Delgado, P. Ezzatti, L. González, Á. Martín, R. Martínez and B. Muracciole, “Proximity tracing applications for Covid-19: data privacy and security,” *XLVII Latin American Computing Conference (CLEI)*, Cartago, Costa Rica, pp. 1-10, 2021.
- [25] A. Lubis and B. Basari, “Proximity-based Covid-19 contact tracing system devices for locally problems solution,” *3rd International Seminar on Research of Information Technology and Intelligent Systems (ISRITI)*, Yogyakarta, Indonesia, pp. 365-370, 2020.
- [26] I. Ozcelik, “CAPEN: cryptographic accumulator based privacy preserving exposure notification,” *9th International Symposium on Digital Forensics and Security (ISDFS)*, Elazig, Turkey, pp. 1-6, 2021.
- [27] H. Xu, L. Zhang, O. Onireti, Y. Fang, W. Buchanan and M. Imran, “BeepTrace: blockchain-enabled privacy-preserving contact tracing for Covid-19 pandemic and beyond,” *IEEE Internet of Things Journal*, vol. 8, no. 5, pp. 3915-3929, 2021.
- [28] L. Garg, E. Chukwu, N. Nasser, C. Chakraborty and G. Garg, “Anonymity preserving IoT-based Covid-19 and other infectious disease contact tracing model,” *IEEE Access*, vol. 8, pp. 159402-159414, 2020.
- [29] S. Sharma, G. Singh, R. Sharma, P. Jones, S. Kraus and Y. Dwivedi, “Digital health innovation: exploring adoption of Covid-19 digital contact tracing apps,” *IEEE Transactions on Engineering Management (Early Access)*, pp. 1-17, 2020.
- [30] V. Shubina, A. Ometov and E. Lohan, “Technical perspectives of contact-tracing applications on wearables for Covid-19 control,” *12th International Congress on Ultra Modern Telecommunications and Control Systems and Workshops (ICUMT)*, Brno, Czech Republic, pp. 229-235, 2020.
- [31] A. Sarkar and S. Ray, “A data driven decision making and contract tracing app for organizations to combat Covid-19,” *International Conference on Computing, Networking, Telecommunications & Engineering Sciences Applications (CoNTESA)*, Tirana, Albania, pp. 88-93, 2020.
- [32] A. de Araujo, I. Garcia, N. Cacho, L. Nascimento, D. Rolim, J. Medeiros, S. Santana, A. Paiva, M. Lima, T. Ramos, K. Macedo, J. Pereira, J. Nascimento, L. Monteiro, M. Ferna. N. Fernandes and F. Lopes, “A platform for citizen cooperation during the Covid-19 pandemic in RN, Brazil,” *IEEE International Smart Cities Conference (ISC2)*, Piscataway, NJ, USA, pp. 1-8, 2020.
- [33] M. Winter, H. Baumeister, U. Frick, M. Tallon, M. Reichert and R. Pryss, “Exploring the usability of the German Covid-19 contact tracing app in a combined eye tracking and retrospective think aloud study,” *43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, Mexico, pp. 2215-2221, 2021.
- [34] M. Sheeraz, A. Athar, A. Hussain, S. Aich, M. Joo and H. Kim, “Blockchain, AI & IoT based Covid-19 contact tracing and distancing framework,” *International Conference on Robotics and Automation in Industry (ICRAI)*, Rawalpindi, Pakistan, pp. 1-6, 2021.
- [35] S. Hsaini, H. Bihri, S. Azzouzi and M. Charaf, “Contact-tracing approaches to fight Covid-19 pandemic: limits and ethical challenges,” *IEEE 2nd International Conference on Electronics, Control, Optimization and Computer Science (ICECOCS)*, Kenitra, Morocco, pp. 1-5, 2020.
- [36] S. Alanoca, N. Jeanrenaud, I. Ferrari, N. Weinberg, R. Çetin, and N. Mialhe, “Digital contact tracing against Covid-19: a governance framework to build trust,” *International Data Privacy Law*, vol. 11, no. 1, pp. 3-17, 2021.
- [37] L. White and P. Basshuysen, “Privacy versus public health? a reassessment of centralised and decentralised digital contact tracing,” *Science and Engineering Ethics*, vol. 27, no. 2, pp. 1-13, 2021.
- [38] S. Vaudenay. (May, 2020). Centralized or decentralized? the contact tracing dilemma. Available: <https://eprint.iacr.org/2020/531>.
- [39] L. Nurgalieva, S. Ryan and G. Doherty, “Attitudes towards Covid-19 contact tracing apps: a cross-national survey,” *IEEE Access (Early Access)*, pp. 1-29, 2021.
- [40] V. Garousi, D. Cutting and M. Felderer, “What do users think of Covid-19 contact-tracing apps? an analysis of eight European apps,” *IEEE Software (Early Access)*, pp. 1-9, 2021.
- [41] S. Tahir, H. Tahir, A. Sajjad, M. Rajarajan and F. Khan, “Privacy-preserving Covid-19 contact tracing using blockchain,” *Journal of Communications and Networks*, vol. 23, no. 5, pp. 360-373, 2021.
- [42] J. Khan, K. Bangalore and K. Ozbay, “COVERT-blockchain: privacy-aware contact tracing for Covid-19 on a distributed ledger,” *3rd Conference on Blockchain Research & Applications for Innovative Networks and Services (BRAINS)*, Paris, France, pp. 31-32, 2021.
- [43] M. Ružička, M. Vološin, J. Gazda and T. Maksymyuk, “Deep learning-based blockchain framework for the Covid-19 spread monitoring,” *International Conference on Electrical, Computer, Communications and Mechatronics Engineering (ICECCME)*, Mauritius, Mauritius, pp. 1-6, 2021.
- [44] E. Bandara, X. Liang, P. Foytik, S. Shetty, C. Hall, D. Bowden, N. Ranasinghe and K. De Zoysa, “A blockchain empowered and privacy preserving digital contact tracing platform,” *Information Processing & Management*, vol. 58, no. 4, pp. 1-17, 2021.
- [45] N. Bari, U. Qamar and A. Khalid, “Efficient contact tracing for pandemics using blockchain,” *Informatics in Medicine Unlocked*, vol. 26, pp. 1-12, 2021.
- [46] T. Mohamed, E. Goda, V. Snasel and A. Hassanien, “Covid-19 contact tracing and detection-based on blockchain technology,” *Informatics*, vol. 8, no. 4, pp. 1-24, 2021.
- [47] D. Siddarth, B. Cantrell, L. Tretikov, P. Eckersley, J. Langford, S. Leibrand, S. Kakade, S. Latta, D. Lewis, S. Tessaro and G. Weyl, “Outpacing the virus: digital response to containing the spread of Covid-

- 19 while mitigating privacy risks,” Edmond J. Safra Center for Ethics, White Paper 5, pp. 1-38, 2020.
- [48] R. Sun, W. Wang, M. Xue, G. Tyson, S. Camtepe and D. Ranasinghe, “An empirical assessment of global Covid-19 contact tracing applications,” IEEE/ACM 43rd International Conference on Software Engineering (ICSE), Madrid, ES, pp. 1085-1097, 2021.
- [49] A. Chen and K. Thio, “Exploring the drivers and barriers to uptake for digital contact tracing,” Social Sciences & Humanities Open, vol. 4, no. 1, pp. 1-13, 2021.
- [50] M. Shahroz, F. Ahmad, M. Younis, N. Ahmad, M. Boulos, R. Vinuesa and J. Qadir, “Covid-19 digital contact tracing applications and techniques: A review post initial deployments,” Transportation Engineering, vol. 5, pp. 1-9, 2021.
- [51] P. Ng, P. Spachos and K. Plataniotis, “Covid-19 and your smartphone: BLE-based smart contact tracing,” IEEE Systems Journal, vol. 15, no. 4, pp. 5367-5378, 2021.
- [52] M. Mancastroppa, C. Castellano, A. Vezzani and R. Burioni, “Stochastic sampling effects favor manual over digital contact tracing,” Nature Communications, vol. 12, no. 1, pp. 1-9, 2021.
- [53] A. Barrat, C. Cattuto, M. Kivelä, S. Lehmann and J. Saramäki, “Effect of manual and digital contact tracing on Covid-19 outbreaks: a study on empirical contact data,” Journal of the Royal Society Interface, vol. 18, no. 178, pp. 20201000, 2020.

Social Customer Relationship Management as a Communication Tool for Academic Communities in Higher Education Institutions through Social Media

Ali Ibrahim¹, Ermatita², Saparudin³

Department of Engineering, Faculty of Engineering, Universitas Sriwijaya, Palembang, Indonesia¹

Department of Information Systems, Faculty of Computer Science, Universitas Sriwijaya, Palembang, Indonesia¹

Multimedia and Game Programming Laboratory, Faculty of Computer Science, Universitas Sriwijaya, Indonesia¹

Department of Information Systems, Faculty of Computer Science, Universitas Sriwijaya, Palembang, Indonesia²

Department of Informatics Engineering, Faculty of Computer Science, Telkom University, Bandung Indonesia³

Abstract—The interaction between academic community members in universities is fundamental in facilitating the learning management process and workflow, thereby necessitating a communication tool system that is unrestricted by time and place. Social customer relationship management as a communication tool through social media is also essential. Therefore, this research employed a mixed-method exploratory design to evaluate 2421 subjects determined based on a proportional random sampling technique from the academic community of universities in four cities in Indonesia. The data collection method used an online interview questionnaire, and the coding techniques, namely open, axial, and selective coding, as well as the value stream analysis, were used. This research found that Social Customer Relationship Management (SCRM) can be a communication tool between the academic communities in higher education. This tool can be exploited by utilizing effective social media platforms, namely Instagram, Facebook, WhatsApp, and YouTube, integrated and connected as a center for information retrieval.

Keywords—Social customer relationship management; communication tools; social media; academic client and communities

I. INTRODUCTION

The implementation of the academic community in a good university is influenced by communication from social media or Social Customer Relationship Management (SCRM) to create a committed and mutually beneficial two-way relationship between customers and the academic community. Because social media has great potential to facilitate closeness between the academic community, allows interaction with many people and customers due to the sophistication of internet-based communication that is increasingly easy and directed, facilitates interaction between users who gain value, and social media is becoming ubiquitous and essential, but this can limit the word-of-mouth interaction of marketers.

Social media that is used as a communication tool has two benefits in human life, the benefits of light and dark sides. Research conducted by Sands et al. (2019) [1] reveals the two sides of social media as a strategy for users. First, social media has a dark side: the identified risks are social media risks for individuals, communities, and organizations. Presence and

Identity load on the same factor. Presence represents concerns about the privacy of the user's presence and activity, whereas Identity reflects concerns about other people or social media entities knowing too much about the user. Apart from that, the bright side of social media is sharing, conversation, relationships, groups, reputation, and personal.

The existence of a debate between SCRM and social media, with the existence of social media, can help communication between the academic community and the students of the University. However, there are technical barriers to social media caused by the lack of facilities and the critical role in the communication process. The research from Jędrzejczyk [2], a weakness in secondary school activities is the absence of a school development strategy that will set guidelines for creating a positive image of the school through the use of social media. The school does not have an active standard-setting on social media. Their activity is limited in terms of monitoring content about them.

As Hujran et al. researched [3], social media platforms have changed how the public sector relates and interacts with its citizens. Indeed, social media offers various benefits to the public sector. However, the use of social media cannot be separated from the risk. Risks that may arise from using social media platforms in the public sector include privacy, security, and lack of control over communication channels.

The academic community in a university consists of lecturers, students, and all campus management bodies, and its involvement, along with the alumni as a supporting group, largely determines the success of a university. Students are the foundation for future environmental protection and the essence of citizens [4]. Therefore, implementers, processes, management, budgets, facilities, infrastructure, testing, evaluation, and related results are necessary to realize good academic quality. Nasution [5] stated that academic quality and resources can be measured based on scientific papers published internationally and globally and indexed by legitimate and recognized databases.

Meanwhile, a lecturer gives class lessons, provides guidance, and updates teaching materials according to scientific developments, thereby allowing students who

eventually become alumni to experience these processes for a particular time. Besides participating in classroom teaching and learning, students must do appropriate scientific work at the end of their studies. Based on their scholarship level and is either a Bachelor's, Master's, or Doctoral program, this work must ensure that the students understand and practice scientific principles and ethics [5].

According to the ASEAN Economic Community (MEA), graduation competencies must be considered to ensure comparability with university alumni domestically and abroad [5]. The ability to compete with local or international alumni depends on the execution of the student's educational process, from teaching to writing scientific papers, and supervisors without this competency will potentially produce incompetent graduates. Hence, members of the academic community, particularly lecturers or teaching staff, must become independent in conducting scientific publications internationally.

Generally, implementing a good university is inseparable from the communication process, which occasionally faces obstacles or difficulties. The term 'communication' comes from the Latin word 'communis,' which means creating togetherness or building friendship between two or more people [6]. According to Harold D. Lasswell, it is a process of delivering messages from a communicator to the communicant through the media to produce specific effects [7].

Communication is a prerequisite of human life and necessary for forming humans, groups, and organizations. Fajar [8] also stated that it aims to change behavior, opinions, attitudes, and social change. However, the university communication process occasionally experiences three main obstacles or difficulties, namely technical, semantic and behavioral barriers that need to be resolved. Technical barriers are caused by several factors, such as the lack of facilities and the necessary roles in the communication process. Social media tools that continue to develop in this field only focus on academic progress and social learning aspects of a vital tool in higher education [9]. Information can be delivered quickly with technology [10], as the sophistication of increasingly accessible and undirected internet-based communication facilitates interaction between users who gain value [11].

Social media became ubiquitous and essential [12] and eliminated the limited word-of-mouth interaction of marketers [13], thereby connecting and empowering customers alongside challenging the implementation of business on the internet. Bala et al. [14] also stated that it serves as a tool for users to interact and exchange information through online communities. Implementing CRM in a college environment includes a student-centered focus, improved customer data and process management, and increased student loyalty, retention, and satisfaction with college programs and services [15]. Social media has great potential to facilitate closeness between academic communities and customers [16] and enables interaction with many people [17]. One of which is a social media-based platform; every university has a different perception of social media-based platforms [18]. A new practice, Social Customer Relationship Management (SCRM), combines the two main concepts between social media or Web

2.0 and Traditional CRM [18], [19]. Faase et al. [20] also concurred that social CRM uses Web 2.0 services to create a committed and mutually beneficial two-way relationship between customers and the academic community.

The transformation of traditional methods into social CRM can accommodate communication between academic community members to achieve common goals. It can enable an organization or community to have real-time conversations with customers, monitor their expressions, and facilitate constructive dialogues [21]. Hence, this strategy is indispensable as a new business approach that can expand the capabilities of today's traditional CRM.

Attaining exemplary university achievements depends on the communication between lecturers, students, education staff, alumni, and other employees who can develop a creative and dynamic spirit [22]. It can trigger an increase in more meaningful interactive relationships and improve aspects of the traditional and social CRM [23]. Online communication can be seen from a person's personality; for example, extroverts convey happiness, optimism, and passion, especially about those that fascinate them. This story will appeal to people with high Openness to experience because these people are curious, open to unusual ideas, and responsive to new things [24]. Social CRM is also regarded as a customer relationship management process that provides communication through social media sites, such as Facebook, Instagram, WhatsApp, and Twitter [25].

According to the rationale above, this research aims to create the basic concept of developing Social Customer Relationship Management (SCRM) as a communication tool through social media for academic communities at universities. Social media is a platform for people to discuss their issues and opinions. Therefore, the process is a breakthrough to improve business achievement through the five domains involved, namely online community building, proactive management of interactions in social communities, online community as a method of customer engagement, the utilization of social media in CRM systems, and collection, as well as the integration and utilization of customer information [26]. According to the rationale above, this research aims to create the basic concept of developing Social Customer Relationship Management (SCRM) as a communication tool through social media for academic communities at universities. Social media is a platform for people to discuss their issues and opinions.

Before knowing the aspects of social media, people must know what social media is? Social media are computer tools that allow people to share or exchange information, ideas, images, videos, and even more with each other through a particular network. This paper covers all aspects of social media with its positive and negative effects. Focus is on particular fields like business, education, society, and youth. This paper describes how these media will broadly affect society [27]. The growing popularity of social media compelled marketers to think about this media and traditional functional marketing areas. Social media is based primarily on the internet or cellular phone based applications and tools to share information. Many social media users are more than some countries' populations today. The impact of social media

on marketing can be judged by comparing marketing before and after the introduction of social media and the type of technologies used in social media [28]. The authors also propose a hypothesis in this study, i.e., “can the social media be used as a communication tool for the academic community in Universities.”

II. BACKGROUND

The Theme of social customer relationship management has been widely studied. From the results of the author's search through the connected paper's platform related to articles on social customer relationship management, the graphic results are obtained as follows in Fig. 1.

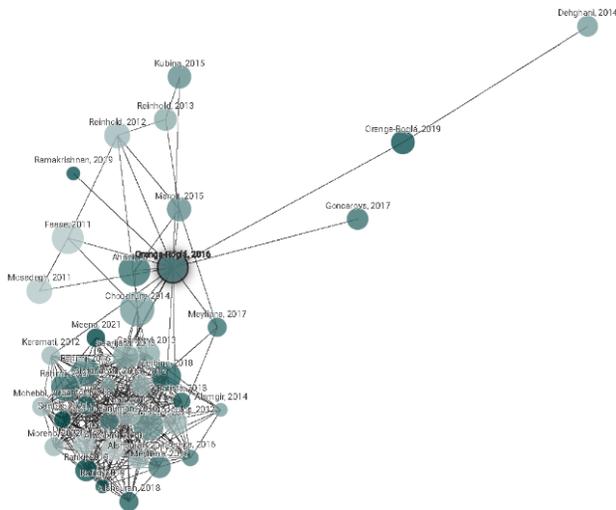


Fig. 1. Social Customer Relationship Management Research Theme Distribution.

Source : <https://bit.ly/3lwcRtR>

From the search results, the theme of social customer relationship (SCRM) has become a research theme that is quite interested in the modern era. Before the SCRM theme emerged, the theme related to Customer Relationship Management had been studied by several previous researchers starting around the years 2010-2014, until entering 2015 the new SCRM theme emerged. However, research related to SCRM only emerged around 2015 when research conducted by Marolt [29] titled “Social CRM Adoption and its Impact on Performance Outcomes:a Literature Review” was published journal *Organizacija* Vol 48 No 4 2015.

Entering 2015 to 2021, the theme of SCRM is increasingly emerging, which can be used as a basis for related research. The author highlights the development of SCRM research themes related to communication in an organization. Research conducted by Dewanarain [30] highlights that using SCRM as a communication tool in hotels can use social media tools for customer relationship management to increase business profitability. In addition, research conducted by Al-Azzam [31] reviews the context of SCRM aimed at improving customer relationships and making customers more engaged.

Research conducted by Arora concludes that a properly implemented SCRM will result in Customer Loyalty, Customer

Retention, and Customer Satisfaction [32]. Dewanarain et al. revealed that social customer relationship management activities could trigger service process innovation in their research, leading to customer engagement with hotel brands [33].

From several previous studies related to the SCRM theme, researchers have tried to link SCRM with customer communication. However, there has been no research that tries to link SCRM as a communication tool for the academic community in higher education through social media. So that the novelty in this research is the development of previous research with the theme of SCRM as a communication tool using social media, so this research is worth doing.

III. LITERATURE REVIEW

Customer relationship management is a strategic business approach underpinned by relationship marketing theory. It is defined as a “process of acquiring, retaining, and partnering with selective customers to create superior value for the company and the customer” [34]. In 2008, CRM underwent a significant shift from a strategy that focused solely on customer transactions to one that integrates customer interactions [35]. Marketers can extract first-hand information about customers, which companies then use to achieve greater effectiveness in delivering customer value [15], [17]. Consequently, CRM was renamed social customer relationship management or CRM 2.0 [35].

Paul Greenberg defined social customer relationship management (SCRM) as a “business strategy of engaging customers through social media with the goal of building trust and brand loyalty” [36]. The introduction of social media has been very disruptive to the customer–marketer relationship; this has raised some speculations about applying traditional CRM models and theoretical concepts [37]. As pointed out by Vivek et al. (2012) and Berthon, Pitt, Plangger, and Shapiro (2012), the traditional approach entailed the firm producing value for customers. However, one of the significant effects garnered by social media is the involvement of customers in the process of value co-creation, either through reviews or in the form of user-generated content [38], [39].

Potential CRM must be applied in the e-learning domain. In this regard, the use of CRM systems is to maximize the value offered by e-learning institutions to students, thereby focusing on managing the institution–student relationship and supporting collaborative relationships with and between scholars and student workgroups [21]. CRM is a business strategy that involves customers and builds trust and brand loyalty [22]. It is defined as a philosophy and strategy where technology platforms, social media, and business rules create an open, transparent, and collaborative environment with customers to satisfy one's needs, thereby achieving mutual benefits [17]. According to Deepa & Deshmukh [23] and Greenberg [17], social CRM is an amalgamation of social media technologies, such as Facebook, Twitter, LinkedIn, YouTube, Pinterest, and other well-known applications, connecting marketers and customers using two-way interaction.

Clear business goals are inseparable from challenges that can undermine the main objective. Any technology operated on

the internet will have significant vulnerabilities regarding data security, legal risks, privacy, and even reputation. Therefore, convincing customers to engage and convert them into business promoters will be more beneficial. It allows the business world to interact effectively with customers, reducing risks and improving all aspects of customer relationships [24].

The growth of social media has significant consequences for the business. Social media have brought in a value-creation environment and information-rich and empowered customers [40]. User-generated content can influence other customers' decisions considerably [41]. Put, social networking sites (SNSs) are changing business activities. They change and reverse how consumers gather information in the decision-making process [42]. Social networking tools give firms a way to track customer feedback but also to proactively respond to customer satisfaction [43], [44].

An example of SCRM in action is seen when airlines provide customers with flight information updates using Twitter or answer customer questions via a Facebook page or when large computer manufacturers like Dell leverage SCRM to drive sales promotions. Thus, social CRM can help companies create a positive customer experience that will help develop customer loyalty and advocacy. Customers will see themselves as partners and feel that they share in the company's success. In the world of higher education, SCRM can give universities the chance to better connect with students who are accustomed to using social platforms since they are primarily from the millennial generation and grew up in the era of web 2.0 [45].

As higher education flocks to marketing, SCRM becomes an acronym rolling to the tongues of these marketing staffers. The lifetime of a university constituent is just as important and essential as that of a product consumer. At its simplest, the high-level goals of SCRM in corporations are to find new customers and maintain existing ones so they can repeat their purchases. Similar goals can be drawn in higher education: to find new students, research funders, faculty, etc., along with retaining them for an extended period of time. Retention in a higher education institution is more of being entrenched in a constituent's life and, hopefully, their legacy [9].

The use of SCRM allows the university to conduct frequent surveys to measure the students' satisfaction, allowing the university to react immediately to student demands – increasing student retention, which is of significant financial value to HEI management [46]. Thus, effective adoption and use of SCRM are of increasing importance to the running of the university [47].

IV. MATERIALS AND METHODS

This study employed a mixed-method exploratory design of qualitative research in the first stage and continued quantitative research, which develops qualitative data. Qualitative research, namely the method, can be understood as a research procedure, and qualitative research, namely the method, can be understood as a research procedure that utilizes descriptive data in the form of written or spoken words from people and actors who can be observed. Qualitative research is conducted to explain and

analyze phenomena, events, social dynamics, attitudes, beliefs, and perceptions of a person or group towards something.

Thus, qualitative research begins by developing basic assumptions and rules of thought that will be used in the research. Qualitative data, including audio recordings and transcripts of in-depth or semi-structured interviews and structured interview questionnaires, contained substantial open-ended comments, including a large number of responses to open-ended comment items [28]. This study employed a mixed-method exploratory design of qualitative research in the first stage and continued quantitative research, which develops qualitative data. The exploratory method is a research method that tends to be carried out to study a problem that has not been clearly defined and provides a good understanding of the research problem. This research method generates ideas or hypotheses for further quantitative research [48].

Qualitative research is a type of research that explores and understands the meaning of several individuals or groups of people originating from social problems. Quantitative research is a type of research that connects variables by requiring researchers to explain how variables affect other variables [49] and the respondents were members of university academic communities, particularly students, lecturers, education staff, and alumni, spread across four provinces in Indonesia, namely South Sumatra, Lampung, Bengkulu, Bangka Belitung. The problem in this research is how to promote higher education to the point where it reaches the intended target, namely students who will enter college through social media. In this study, it can be concluded that this research uses qualitative research methods and then continues with quantitative research, aiming that SCRM in universities can be used as a communication tool for the academic community. The following is a description of the flow in this research in Fig. 2.



Fig. 2. Research Flow.

The respondents were determined through John Creswell's technique [50], which suggested 350 subjects for survey research using a proportional random sampling method. Then, the data were collected through an online interview questionnaire on the Google Form platform. It was arranged based on open and closed interview techniques, which respectively allowed the expression of respondents through individual response sentences and provided answers to be chosen [50]. A structured interview guide was used separately from the research questions in this study because this study used two methods at once. Thus, structured interview guides were used in a qualitative setting, while mixed questions were used in a quantitative setting.

This research also employed a coding technique consisting of open, axial, and selective coding stages. Open coding is a process of breaking down data, studying it one by one, comparing, and conceptualizing data. So the goal at this stage is to name and categorize phenomena through data scrutiny. Axial coding is a procedure where data is put back in a new way after open coding by making connections between categories. Selective coding is selecting core categories,

connecting them with other categories systematically, validating these relationships, and filling in categories that require refinement and development [49]. Also, the descriptive data obtained from the questionnaires were analyzed in in-depth analysis to determine the main themes of the problem, using the ATLAS program: The Qualitative Data Analysis & Research Software and the Microsoft Excel program for classifying [51]. In addition, the value stream analysis was employed to determine the effectiveness of SCRM modeling as a tool for interaction in the academic community through social media. The sampling was done through qualitative and quantitative methods. The method used for qualitative was using the exploratory method, then quantitatively using the method used was proportional random sampling from John Creswell. Data collection was also through a Google form interview questionnaire, and the technique of the questionnaire was carried out openly and closed. Data processing using ATLAS program assistance: analysis and research software then also using Microsoft office excel applications and value stream analysis [51].

V. RESULTS

A. Research Subject Demographic Data

The total research subjects were $N = 2421$ members of the academic community in universities and consisted of students ($N = 2141/88.43\%$), lecturers ($92/3.8\%$), education staff ($23/0.95\%$), and alumni ($165/6.82\%$). Furthermore, the male subjects were $N = 1449/60\%$, while the females were $N = 972/40\%$.

B. Ownership of Social Media Accounts in the Academic Community of the Universities

In this research, a survey was conducted regarding the ownership of social media accounts in the university community. Based on the results, the most popular platform is WhatsApp ($N = 2416$), followed by Facebook ($N = 1374$), Instagram ($N = 765$), Twitter ($N = 457$), YouTube ($N = 319$), Pinterest ($N = 250$), LinkedIn ($N = 134$), MySpace ($N = 64$), and Flickr ($N = 24$).

C. Social Media Visited by the Academic Community to obtain University Information

There were nine types of social media used to obtain university information, with the highest to the lowest order being Instagram ($N = 1783$), Facebook ($N = 946$), YouTube ($N = 869$), WhatsApp ($N = 509$), Twitter ($N = 159$), LinkedIn ($N = 156$), Pinterest ($N = 126$), MySpace ($N = 58$), and Flickr ($N = 17$). On further analysis, the three highest social media platforms used were Instagram ($N = 1783$), Facebook ($N = 946$), YouTube ($N = 869$), while the three lowest were LinkedIn ($N = 156$), Pinterest ($N = 126$), MySpace ($N = 58$), and Flickr ($N = 17$).

D. Social Media Utilization as a Tool for Interaction of the Academic Community

According to the analysis results of the social media utilization as a means of interaction for the academic community in universities, the total number of subjects was 2421. From this number, 1367 answered that they actively interacted on social media, 744 responded that it was

somewhat ineffective, 200 were inactive, and 110 stated they had not implemented social media in interactions. A graphic representation of the data distribution of the interaction of university communities is presented in Fig. 3.

E. Information Sought by the Academic Community through Social Media

According to the survey results, the most searched information on social media is lecture information ($N = 1368$). This is followed by information on requirements to become a student ($N = 1285$), tuition fees ($N = 1268$), research program accreditation ($N = 1174$), university accreditation ($N = 1114$), student activities ($N = 964$), facilities ($N = 846$), location of the university ($N = 715$), staffing ($N = 417$), alumni ($N = 360$), and cooperation links ($N = 292$).

The three most sought information by students were lecture information ($N = 1221$), requirements to become a student ($N = 1203$), and tuition fees ($N = 1159$), while lecturers majorly searched for information on student activities ($N = 50$), facilities ($N = 49$), and lectures ($N = 47$). Conversely, education staff looked for information on employment ($N = 10$), lectures and facilities ($N = 18$), alongside accreditation of research programs and the university ($N = 16$), while alumni often searched for information on lectures ($N = 91$), research program accreditation ($N = 83$), and university accreditation ($N = 80$).

F. University Marketing through Social Media

The survey regarding the assessment of university marketing by the academic community ($N = 2421$) discovered: 1) active implementation ($N = 1817$), 2) less active ($N = 522$), 3) inactive ($N = 53$), and 4) respondents who have not implemented social media ($N = 29$). A graphic illustration of the assessment of the university marketing strategy is presented in Fig. 4.

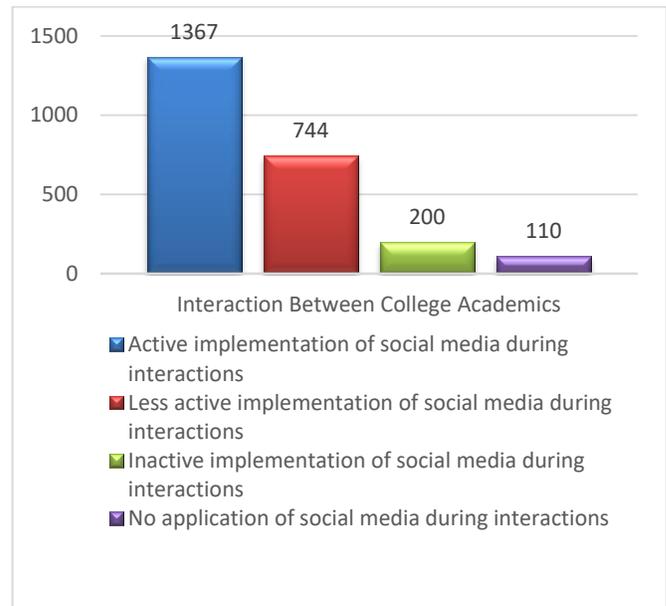


Fig. 3. Interaction between University Academics.

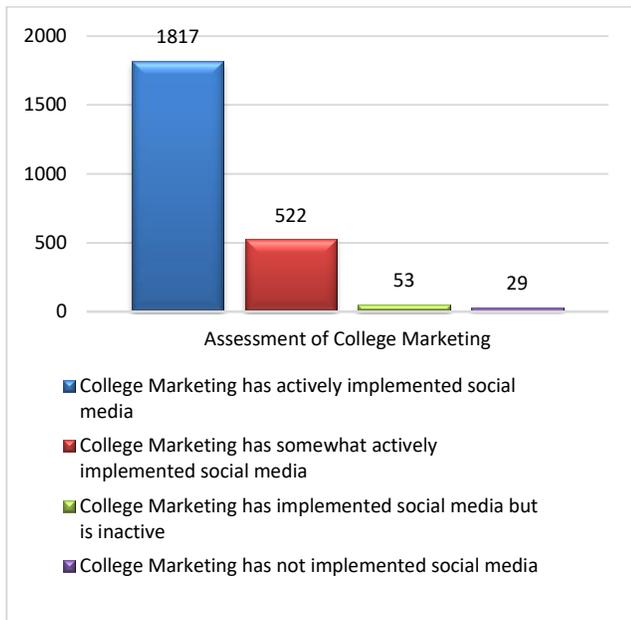


Fig. 4. Assessment of University Marketing.

G. Social Media as a Communication Tool in Universities

To find out whether the hypothesis proposed in this study is proven. The author conducts an analysis using the Value Stream Analysis technique to see how the relationship between social media as a communication tool in universities is. From the results of a survey conducted by four types of respondents in universities, namely students, lecturers, education staff, and alumni, stated that the interaction between the academic community in universities in terms of using social media, the results are as follows, in Fig. 5.

From the survey results, it can be seen that from N = 2421 that most of the interactions between the academic community in universities have used social media, only the implementation is still not the same in every university. Several social media platforms were recommended to help the higher education academic community, namely Instagram (N = 1929), followed by Facebook (N = 1149), WhatsApp (N = 1127), YouTube (N = 1040), Twitter (N = 329), LinkedIn (N = 161), Pinterest (N = 155), MySpace (N = 91), and Flickr (N = 43). Some platforms were also recommended for searching the information, namely Instagram (N = 1783), Facebook (N = 946), YouTube (N = 869), and WhatsApp (N = 509).

Hence, the four highest social media platforms recommended as communication tools in academic communities were Instagram (N = 1929), Facebook (N = 1149), WhatsApp (N = 1127), and YouTube (N = 1040). The reasons for this proposal were based on the findings that social media promoted: 1) ease of access (n = 1525), 2) complete information (1236), 3) updated information (1181), and 4) attractive appearance (704). Fig. 6 shows an infographic representation of social media recommended by the academic community in higher education.

An analysis to determine the modeling of Social Customer Relationship Management (SCRM) as a communication tool in universities through social media was conducted based on the

research findings and prepared via the value stream analysis technique. The results of this study are Social Customer Relationship Management (SCRM) as a communication tool for the academic community, and targeted promotions for universities. It can be seen that social media is often encountered to become a communication tool used by the academic community. One of them is that each university has a different perception of social media-based platforms.

That facilitates communication between academic community members through social media platforms that aim to enable information provision and acquisition. Based on research findings, social media platforms that allow the development of SCRM as a communication tool between members of the academic community are Instagram, Facebook, WhatsApp, and YouTube. These four platforms can enable connections and relationships alongside providing and conveying information, leading to their recommendation by academic communities as university communication tools. An overview of the connectedness of SCRM as a communication tool through social media in the value stream analysis is presented in Fig. 7.

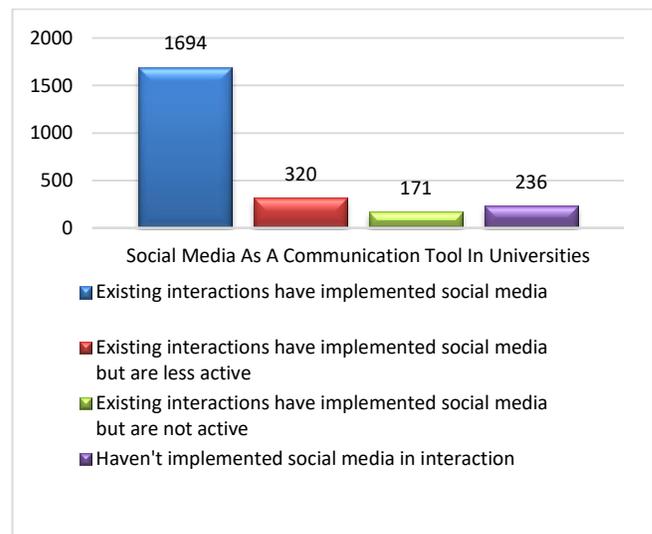


Fig. 5. Social Media as a Communication Tool in Universities.

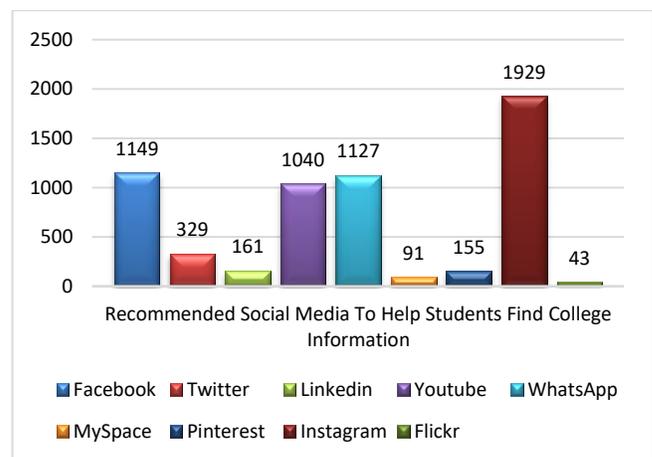


Fig. 6. Recommended Social Media to Help the Academic Community Find University Information.

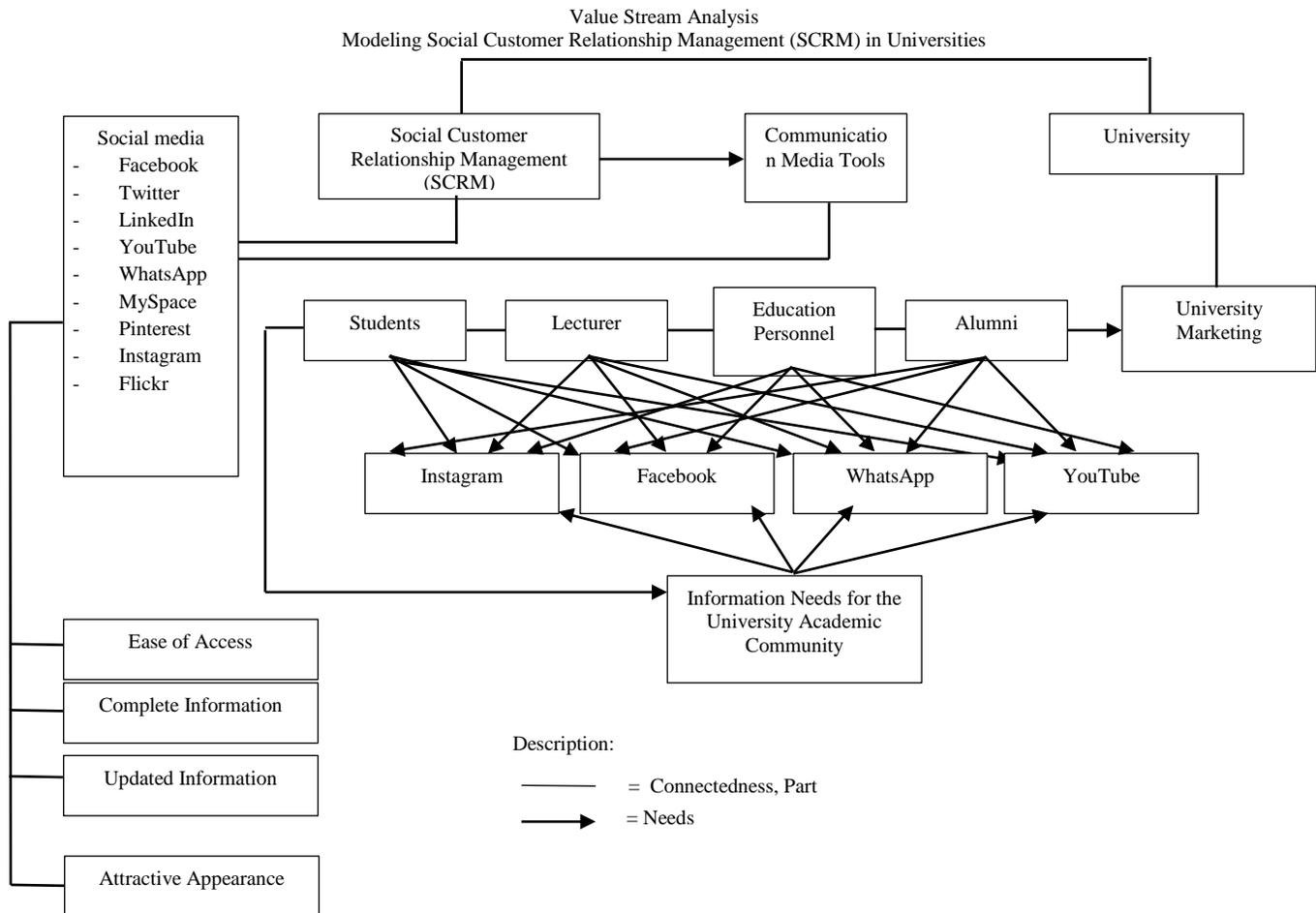


Fig. 7. Modeling Social Customer Relationship Management (SCRM) in Universities.

VI. DISCUSSION

The use of social media has grown and transformed into a communication tool [52], [53], including in modern times in academic communities. Based on survey data, the academic community members mostly used WhatsApp (N = 2416), Facebook (N = 1374), and Instagram (N = 765), while the rest chose other platforms. These three applications are the most used social media platforms in the world.

Meanwhile, the new form of CRM, known as SCRM [28], has become a necessity in universities as a means of communication and interaction between academic community members. Sigala [54] stated that SCRM could provide users with benefits and satisfaction [55]. Interaction between internet users is the core of the social web [56], and through social media, universities can join communities outside their official boundaries to obtain information [54]. It corresponds with Diffley and McCole [57], who stated that social CRM could represent routine activities, including in social networks, to build long-term relationships that involve customers, and create shared value in the development and empowerment [54] of university innovation.

Social media is a means of interaction and communication, as evidenced by its use in the academic community's search for university information in this research. The three most popular

social media platforms used as tools to obtain information were Instagram (N = 1783), Facebook (N = 946), and YouTube (N = 869), which are also some of the most common applications [46]. According to the result presented in Section 3, this study proves the hypothesis that the social media is a potential communication media between the academic communities in higher education.

Furthermore, 1367 (56%) respondents in this research admitted that they employed social media during interactions, while 1054 (54%) gave responses of less active, inactive, and unimplemented interactions. These findings necessitate the creation of a system that regulates social CRM through social media as a means of communication and interaction between members of the academic community in universities.

Social media is defined as a technology used for interaction through the web [58], [59]. Hence, academic communities need to properly engage customers on related platforms [60], [61]. Social media is one of the most effective tools for all kinds of information and has many advantages that allow its users to connect easily over the internet. It can also help students and the academic community in teaching and learning, global collaboration, and provide relevant opportunities in one click [62]. The search and distribution of information through social media by the academic community is presented in Table I.

TABLE I. INFORMATION WANTED BY THE ACADEMIC COMMUNITY

No	Academic Community	Information Sought on Social Media
1	Students	Information on lectures, requirements to become a student, and tuition fees
2	Lecturers	Information on student activities, facilities, and lectures
3	Education Personnel	Information on staffing, lectures, and facilities, alongside research program and university accreditation
4	Alumni	Information on lectures as well as research program and university accreditation

This study also proved that social media is an essential communication tool for the university academic community is very important, as evidenced by the recommendations to use related platforms, namely Instagram (N = 1929), Facebook (N = 1149), WhatsApp (N = 1127), and YouTube (N = 1040), to obtain information universities.

The current generation is probably the loneliest, and this can be tackled by image-based platforms such as Instagram [63]. This application was launched in 2010 through the App Store and reached 2 million users in precisely two months. It has also gained wide global popularity, where one of the recent trends consists of posting side-by-side photos comprising one ideal and realistic depiction [64]. It is highly remarkable compared to the development of other social media platforms, such as Twitter and Foursquare, which took one to two years to reach the same number of users [65]. In addition, Instagram has changed the scope of education because students can share experiences and exchange ideas and solutions with other students, lecturers, and individuals [66].

Meanwhile, Facebook is a platform that has rapidly developed and is the second-largest global social network and the largest in Turkey [67]. It has facilitated the most substantial impact on the intention to use social media [68], [69] and fulfills people's needs to show affection and vent feelings [70]. Facebook is used for making friends, surveillance, videos, photos, music, sharing ideas, games, organizations, politics, commerce, and communication in education [71]. The use of social applications in education helps expand the learning experience and enhance interactions inside and outside the classroom, such as WhatsApp, which fundamental functions as a tool for communication between members of the academic community [72].

WhatsApp is an application used to send messages [71] and complement face-to-face communication [73]. The WhatsApp messenger application is widely used among students and the academic community for sending messages, photos, videos, and audio [74]. According to Cheung et al. [75], WhatsApp is a mobile service that allows the academic community to interact socially, enables learning and solving difficulties, and considers an alternative platform for distributing assignments, ideas, and information. Hence, this platform has become essential to building feelings of unity with people [76].

Furthermore, as the most famous of these sites, YouTube exceeds two billion views per day, with new videos uploaded on an average every minute. Not only as a video content

channel but YouTube is also a new educational tool that attracts public attention [77] and is best understood as a means of disseminating information [78]. This application offers four ways to sort search results: relevance, upload date, number of views, and ratings [79]. It also has stringent quality requirements, such as detecting the packets belonging to the application in the packet stream and determining the appropriate quality parameters [80].

In addition, the assessment of marketing at universities by the academic community in this research showed that 1817 (75%) from the total of 2421 admitted that college marketing had actively implemented social media, proving its importance as an interaction system. Advancing the mission and programs of the academic community is very important to assess the effective use of social media from a public perspective. As stated previously, the public tends to accept and use information conveyed through social media, which is believed to be more credible than other platforms [40]. The use of social media is the key to developing interactions that should involve openness, transparency, usability, and interactive features [81], [82].

Social media should become a communication interaction system in universities. Social media is also referred to as social sensing in universities' academic communities for four reasons: ease of access, updated information, complete information, and attractive appearance. Meanwhile, the benefits are information sharing, and publicity, alongside offering and receiving support and advice [83], [84]. Hence, people in universities can interact freely without fear of adequate fluency in spoken English through social media [85], [86],[87].

VII. CONCLUSION

This study concludes that Social Customer Relationship Management (SCRM) can be a means of communication between the academic community in higher education. This is evident from the graphics contained in this study showing that social media is used as a communication tool. SCRM modeling is a bit difficult for readers to understand and understand from research. From the results, it was found that social media is a critical communication tool for the university's academic community, which is very important, as evidenced by the recommendation to use related platforms, namely, Instagram (N = 1929), Facebook (N = 1149), WhatsApp (N = 1127), and YouTube (N = 1040), for university information. Furthermore, 1367 (56%) respondents in this study admitted to using social media when interacting, while 1054 (54%) gave responses in inactive, inactive, and not implemented interactions. These findings require creating a system that regulates social CRM through social media as a means of communication and interaction between members of the academic community in universities, as seen in Fig. 5.

This research concluded that Social Customer Relationship Management (SCRM) could be a communication tool between academic community members in universities. It can be conducted by utilizing effective social media platforms, such as Instagram, Facebook, WhatsApp, and YouTube, integrated and connected as an information search center. Consequently, social media has become an SCRM tool due to: 1) ease of access, 2) complete information, 3) updated information, and

4) attractive appearance. A structured interview guide was used separately from the research questions in this study because this study used two methods at once. Thus, structured interview guides were used in a qualitative setting, while mixed questions were used in a quantitative setting. This study proves the hypothesis that social media is a potential communication medium between academic communities in higher education.

The author also recommends further research with a purely quantitative approach, such as (experiments, correlations, and surveys) so that the relationship between Social Customer Relationship Management (SCRM) can be seen significantly as a communication tool in the academic community in higher education.

VIII. RECOMMENDATION

Recommendations in this study are by looking at some of the weaknesses of this study. So that it becomes suggestions and input for future research. Such as: 1) Revealing more about technical barriers in the use of social media as an academic communication tool; 2) Using more diverse research methods, such as pure quantitative and other mixed-method approaches; 3) Conducting further research with survey analysis with a quantitative approach.

ACKNOWLEDGMENT

We would like to thank the management of Universitas Sriwijaya, Department of Engineering, Faculty of Engineering Universitas Sriwijaya. Faculty of Computer Science and Multimedia and Game Programming Laboratory, Universitas Sriwijaya, Indonesia, for supporting this research.

REFERENCES

- [1] S. Sands, C. Campbell, C. Ferraro, and A. Mavrommatis, "Seeing light in the dark: Investigating the dark side of social media and user response strategies," *Eur. Manag. J.*, vol. 38, no. 1, pp. 45–53, Feb. 2020, doi: 10.1016/J.EMJ.2019.10.001.
- [2] W. Jedrzejczyk, "Barriers in the use of social media in managing the image of educational institutions," *Procedia Comput. Sci.*, vol. 192, pp. 1904–1913, 2021, doi: 10.1016/J.PROCS.2021.08.196.
- [3] O. Hujran, M. M. Al-Debei, and R. Alhawsawi, "Potential barriers to the use of social media in the public sector: Lessons from Saudi Arabia," *Int. J. Bus. Inf. Syst.*, vol. 36, no. 1, pp. 119–143, 2021, doi: 10.1504/IJBIS.2021.112397.
- [4] F. Tong, C. Lin, R. Intra, K. Anusawari, K. B. Khen, and K. Thep, "Research on the Influence of Mass Communication on College Students' Environmental Behavior from the Perspective of Media Convergence," 2010, doi: 10.1088/1755-1315/576/1/012010.
- [5] M. K. M. Nasution, "Social Network Mining (SNM): A Definition of Relation between the Resources and SNA," *Int. J. Adv. Sci. Eng. Inf. Technol.*, vol. 6, no. 6, pp. 975–981, Dec. 2016, doi: 10.18517/IJASEIT.6.6.1390.
- [6] H. Cangara, *Introduction to Communication Science*, 2nd ed. Jakarta: PT Raja Grafindo Persada, 2014.
- [7] O. U. Effendy, *Communication Science Theory and Practice*. Bandung: PT. Remaja Rosda Karya, 2000.
- [8] M. Fajar, *Communication Science Theory and Practice*. Yogyakarta: Graha Ilmu, 2009.
- [9] V. Farnsworth, "Social Customer Relationship Management in higher education," *Open Access Theses*, Apr. 2016, Accessed: May 11, 2022. [Online]. Available: https://docs.lib.purdue.edu/open_access_theses/768.
- [10] Wursanto, *Organizational Science Fundamentals*. Yogyakarta: Andi, 2005.
- [11] C. T. Carr and R. A. Hayes, "Social Media: Defining, Developing, and Divining," <https://doi.org/10.1080/15456870.2015.972282>, vol. 23, no. 1, pp. 46–65, Jan. 2015, doi: 10.1080/15456870.2015.972282.
- [12] S. Asur and B. A. Huberman, "Predicting the future with social media," *Proc. - 2010 IEEE/WIC/ACM Int. Conf. Web Intell. WI 2010*, vol. 1, pp. 492–499, 2010, doi: 10.1109/WI-IAT.2010.63.
- [13] T. Bergström, L. Bäckman, and S. Ross, "Marketing and PR in Social Media," 2013.
- [14] R. Bala, V. Krishnan, and W. Zhu, "Distributed Development and Product Line Decisions," *Prod. Oper. Manag.*, vol. 23, no. 6, pp. 1057–1066, Jun. 2014, doi: 10.1111/POMS.12185.
- [15] E. D. Seeman and M. O'Hara, "Customer relationship management in higher education: Using information systems to improve the student-school relationship," *Campus-Wide Inf. Syst.*, vol. 23, no. 1, pp. 24–34, 2006, doi: 10.1108/10650740610639714/FULL/XML.
- [16] C. Heller and B. G. Parasnis, "Strategy & Leadership From social media to social customer relationship management," *J. Strateg. Leadersh.*, 2011, doi: 10.1108/10878571111161507.
- [17] M. Dewing, *Social Media: An Introduction*. Ottawa: Library of Parliament, 2010.
- [18] O. Dospinescu and N. Dospinescu, "Perception Over E-Learning Tools In Higher Education: Comparative Study Romania And Moldova," 2020, doi: 10.24818/ie2020.02.01.
- [19] S. Mohan, E. Choi, and D. Min, "Conceptual modeling of enterprise application system using social networking and web 2.0 'social CRM system,'" *Proc. - 2008 Int. Conf. Converg. Hybrid Inf. Technol. ICHIT 2008*, pp. 237–244, 2008, doi: 10.1109/ICHIT.2008.263.
- [20] R. Faase, R. Helms, and M. Spruit, "Web 2.0 in the CRM domain: Defining social CRM," *Int. J. Electron. Cust. Relatsh. Manag.*, vol. 5, no. 1, pp. 1–22, 2011, doi: 10.1504/IJECRM.2011.039797.
- [21] C. Olszak and T. Bartus, "Multi-Agent Framework for Social Customer Relationship Management Systems," *Issues Informing Sci. Inf. Technol.*, vol. 10, 2013.
- [22] N. Vera, "Lecturer and Student Communication Strategy in Improving the Quality of Online Learning During the Covid-19 Pandemic," *Avant Garde*, vol. 8, no. 2, pp. 165–177, Dec. 2020, doi: 10.36080/AG.V8I2.1134.
- [23] B. K. S. Dr. Lokesh Arora, "Social Customer Relationship Management: A Literature Review," *Eurasian J. Anal. Chem.*, vol. Volume 13, no. Issue 6, pp. 308–316, 2018, Accessed: Feb. 08, 2022. [Online]. Available: <http://www.eurasianjournals.com/abstract.php?id=900>.
- [24] B. Anastasiei and N. Dospinescu, "A model of the relationships between the Big Five personality traits and the motivations to deliver word-of-mouth online," *Psihologija*, vol. 51, no. 2, pp. 215–227, 2018, doi: 10.2298/PSI161114006A.
- [25] S. Gaidhani, L. Arora, B. Kumar Sharma, and A. Professor, "Understanding The Attitude Of Generation Z Towards Workplace," *Int. J. Manag. Technol. Eng.*, 2019.
- [26] K. Kantorová and P. Bachmann, "Social Customer Relationship Management and Organizational Characteristics," *Inf.* 2018, Vol. 9, Page 306, vol. 9, no. 12, p. 306, Dec. 2018, doi: 10.3390/INFO9120306.
- [27] S. Siddiqui, "Social Media its Impact with Positive and Negative Aspects," *Int. J. Comput. Appl. Technol. Res.*, vol. 5, no. 2, pp. 71–75, 2016, Accessed: Mar. 25, 2022. [Online]. Available: www.ijcat.com.
- [28] C. Anderson, "Presenting and evaluating qualitative research," *Am. J. Pharm. Educ.*, vol. 74, no. 8, 2010, doi: 10.5688/AJ7408141.
- [29] M. Marolt, A. Pucihar, and H.-D. Zimmermann, "Social CRM Adoption and its Impact on Performance Outcomes: a Literature Review," *Organizacija*, vol. 48, no. 4, pp. 260–271, Dec. 2015, doi: 10.1515/ORG-2015-0022.
- [30] S. Dewnarain and H. Ramkissoon, "Social customer relationship management in the hospitality industry," *J. Hosp.*, vol. 1, no. 1, 2019, [Online]. Available: <https://www.researchgate.net/publication/333755634>.
- [31] A. Al-Azzam and R. T. Khasawneh, "Social Customer Relationship Management (SCRM): A Strategy for Customer Engagement," *IGI Glob.*, pp. 45–58, Nov. 2017, doi: 10.4018/978-1-5225-1686-6.CH003.

- [32] L. Arora, P. Singh, V. Bhatt, and B. Sharma, "Understanding and managing customer engagement through social customer relationship management," <https://doi.org/10.1080/12460125.2021.1881272>, vol. 30, no. 2-3, pp. 215-234, 2021, doi: 10.1080/12460125.2021.1881272.
- [33] S. Dewnarain, H. Ramkissoon, and F. Mavondo, "Social customer relationship management: a customer perspective," *J. Hosp. Mark. Manag.*, vol. 30, no. 6, pp. 673-698, 2021, doi: 10.1080/19368623.2021.1884162.
- [34] A. Parvatiyar and J. N. Sheth, "Customer relationship management: Emerging practice, process, and discipline," *J. Econ. Soc. Res.*, vol. 3, no. 2, 2001, Accessed: May 11, 2022. [Online]. Available: https://www.researchgate.net/publication/312458264_Customer_relationship_management_Emerging_practice_process_and_discipline.
- [35] S. Shokohyar, R. Tavallaee, and K. Keramatnia, "Identifying Effective Indicators in the Assessment of Organizational Readiness for Accepting Social CRM," *Int. J. Manag. Account. Econ.*, vol. 9, no. 4, 2017, Accessed: May 11, 2022. [Online]. Available: https://www.ijmae.com/article_115232.html.
- [36] N. Woodcock, A. Green, and M. Starkey, "Social CRM as a business strategy," *J. Database Mark. Cust. Strateg. Manag.*, vol. 18, no. 1, pp. 50-64, Mar. 2011, doi: 10.1057/DBM.2011.7/FIGURES/5.
- [37] P. Harrigan, G. Soutar, M. Choudhury, and M. Lowe, "Modelling CRM in a social media age," *Australas. Mark. J.*, 2014, doi: 10.1016/j.ausmj.2014.11.001.
- [38] S. D. Vivek, S. E. Beatty, and R. M. Morgan, "Customer Engagement: Exploring Customer Relationships Beyond Purchase," <http://dx.doi.org/10.2753/MTP1069-6679200201>, vol. 20, no. 2, pp. 122-146, Apr. 2014, doi: 10.2753/MTP1069-6679200201.
- [39] P. R. Berthon, L. F. Pitt, K. Plangger, and D. Shapiro, "Marketing meets Web 2.0, social media, and creative consumers: Implications for international marketing strategy," *Bus. Horiz.*, vol. 55, no. 3, pp. 261-271, May 2012, doi: 10.1016/J.BUSHOR.2012.01.007.
- [40] T. Hennig-Thurau et al., "The Impact of New Media on Customer Relationships," <http://dx.doi.org/10.1177/1094670510375460>, vol. 13, no. 3, pp. 311-330, Aug. 2010, doi: 10.1177/1094670510375460.
- [41] G. Viglia, R. Minazzi, and D. Buhalis, "The influence of e-word-of-mouth on hotel occupancy rate," *Int. J. Contemp. Hosp. Manag.*, vol. 28, no. 9, pp. 2035-2051, 2016, doi: 10.1108/IJCHM-05-2015-0238/FULL/XML.
- [42] C. H. Li and C. M. Chang, "The influence of trust and perceived playfulness on the relationship commitment of hospitality online social network-moderating effects of gender," *Int. J. Contemp. Hosp. Manag.*, vol. 28, no. 5, pp. 924-944, 2016, doi: 10.1108/IJCHM-05-2014-0227/FULL/XML.
- [43] K. L. Xie, Z. Zhang, Z. Zhang, A. Singh, and S. K. Lee, "Effects of managerial response on consumer eWOM and hotel performance: Evidence from TripAdvisor," *Int. J. Contemp. Hosp. Manag.*, vol. 28, no. 9, pp. 2013-2034, 2016, doi: 10.1108/IJCHM-06-2015-0290/FULL/XML.
- [44] O. Maecker, C. Barrot, and J. U. Becker, "The effect of social media interactions on customer relationship management," *Bus. Res.*, vol. 9, no. 1, pp. 133-155, Apr. 2016, doi: 10.1007/S40685-016-0027-6/TABLES/11.
- [45] J. Wongsansukcharoen, J. Trimetsontorn, and W. Fongsuwan, "Social customer relationship management and differentiation strategy affecting banking performance effectiveness," *Res. J. Bus. Manag.*, vol. 7, no. 1, pp. 15-27, 2013, doi: 10.3923/RJBM.2013.15.27.
- [46] T. Daradoumis, I. Rodriguez-Ardura, J. Faulin, A. A. Juan, F. Xhafa, and F. J. Martinez-Lopez, "Customer relationship management applied to higher education: Developing an e-monitoring system to improve relationships in electronic learning environments," *Int. J. Serv. Technol. Manag.*, vol. 14, no. 1, pp. 103-125, 2010, doi: 10.1504/IJSTM.2010.032887.
- [47] C. R. Montenegro de Lima, T. Coelho Soares, M. Andrade de Lima, M. Oliveira Veras, and J. B. S. O. de A. Andrade Guerra, "Sustainability funding in higher education: a literature-based review," *Int. J. Sustain. High. Educ.*, vol. 21, no. 3, pp. 441-464, Apr. 2020, doi: 10.1108/IJSHE-07-2019-0229.
- [48] N. Mbaka and O. Monday Isiramen, "The Changing Role Of An Exploratory Research In Modern Organisation," *Int. J. Bus. Manag.*, vol. 4, no. 12, 2021, Accessed: Mar. 25, 2022. [Online]. Available: <http://www.gphjournal.org/index.php/bm/article/view/524>.
- [49] J. W. Creswell, *Research design: qualitative, quantitative, and mixed methods approaches*. Thousand Oaks: SAGE Publications, Inc., 2014.
- [50] J. W. Creswell, *Educational research: planning, conducting, and evaluating quantitative and qualitative research*. Boston: Pearson, 2012.
- [51] ATLAS.ti, "ATLAS.ti 9 Windows - User Manual." Scientific Software Development GmbH, Berlin, 2021.
- [52] C. McClellan, M. M. Ali, R. Mutter, L. Kroutil, and J. Landwehr, "Using social media to monitor mental health discussions - evidence from Twitter," *J. Am. Med. Inform. Assoc.*, vol. 24, no. 3, pp. 496-502, 2017, doi: 10.1093/JAMIA/OCW133.
- [53] K. Wilson and J. Keelan, "Social media and the empowering of opponents of medical technologies: The case of anti-vaccinationism," *J. Med. Internet Res.*, vol. 15, no. 5, 2013, doi: 10.2196/jmir.2409.
- [54] M. Sigala, "Implementing social customer relationship management: A process framework and implications in tourism and hospitality," *Int. J. Contemp. Hosp. Manag.*, vol. 30, no. 7, pp. 2698-2726, Sep. 2018, doi: 10.1108/IJCHM-10-2015-0536.
- [55] S. Orenga-Roglá and R. Chalmeta, "Social customer relationship management: taking advantage of Web 2.0 and Big Data technologies," *Springerplus*, vol. 5, no. 1, pp. 1-17, Dec. 2016, doi: 10.1186/S40064-016-3128-Y/FIGURES/2.
- [56] R. Alt and O. Reinhold, "Social Customer Relationship Management (Social CRM)," *Bus. Inf. Syst. Eng.* 2012 45, vol. 4, no. 5, pp. 287-291, Sep. 2012, doi: 10.1007/S12599-012-0225-5.
- [57] S. Duffley, P. McCole, and E. Carvajal-Trujillo, "Examining social customer relationship management among Irish hotels," *Int. J. Contemp. Hosp. Manag.*, vol. 30, no. 2, pp. 1072-1091, 2018, doi: 10.1108/IJCHM-08-2016-0415/FULL/XML.
- [58] K. Yawied, L. Ellis, and M. Wong, "A framework for the adoption of social customer relationship management (scrm) by private sector," *Asian J. Sci. Technol.*, vol. 9, no. 4, 2018, Accessed: Feb. 08, 2022. [Online]. Available: <https://www.journalajst.com/framework-adoption-social-customer-relationship-management-scrm-byprivate-sector>.
- [59] F. Mazyed, F. Aldaihani, N. Azman, and B. Ali, "Impact of Social Customer Relationship Management on Customer Satisfaction through Customer Empowerment: A Study of Islamic Banks in Kuwait," *Int. Res. J. Financ. Econ.*, 2018, Accessed: Feb. 08, 2022. [Online]. Available: <http://www.internationalresearchjournaloffinanceandeconomic.com>.
- [60] Choi and Theoni, "Social media marketing must start at the top: Companies need to make their efforts more of an 'inside job,'" *Strateg. Dir.*, vol. 32, no. 5, pp. 25-27, 2016, doi: 10.1108/SD-02-2016-0023/FULL/XML.
- [61] F. Li, J. Larimo, and L. C. Leonidou, "Social media marketing strategy: definition, conceptualization, taxonomy, validation, and future agenda," *J. Acad. Mark. Sci.* 2020 491, vol. 49, no. 1, pp. 51-70, Jun. 2020, doi: 10.1007/S11747-020-00733-3.
- [62] M. R. Kalasi, "The Impact of Social Networking on New Age Teaching and Learning: An Overview," *J. Educ. Soc. Policy*, vol. 1, no. 1, 2014, Accessed: Feb. 08, 2022. [Online]. Available: www.jespnet.com.
- [63] M. Pittman and B. Reich, "Social media and loneliness: Why an Instagram picture may be worth more than a thousand Twitter words," 2016, doi: 10.1016/j.chb.2016.03.084.
- [64] M. Tiggemann and I. Anderberg, "Social media is not real: The effect of 'Instagram vs reality' images on women's social comparison and body image," <https://doi.org/10.1177/1461444819888720>, vol. 22, no. 12, pp. 2183-2199, Nov. 2019, doi: 10.1177/1461444819888720.
- [65] Z. A. Latiff and N. A. S. Safiee, "New Business Set Up for Branding Strategies on Social Media - Instagram," *Procedia Comput. Sci.*, vol. 72, pp. 13-23, Jan. 2015, doi: 10.1016/J.PROCS.2015.12.100.
- [66] J. P. Carpenter, S. A. Morrison, M. Craft, and M. Lee, "How and why are educators using Instagram?," *Teach. Teach. Educ.*, vol. 96, p. 103149, Nov. 2020, doi: 10.1016/J.TATE.2020.103149.

- [67] D. N. H. SÖYLEMEZ and P. D. B. ORAL, "Student Teachers' Perceptions of Multicultural Education Based on Their Social Media Use," *J. Educ. Cult. Stud.*, vol. 4, no. 1, p. 103, Feb. 2020, doi: 10.22158/JECS.V4N1P103.
- [68] A. Alarabiat and S. Al-Mohammad, "The potential for Facebook application in undergraduate learning: A study of Jordanian students," *Interdiscip. J. Information, Knowledge, Manag.*, vol. 10, pp. 81–103, 2015, Accessed: Feb. 08, 2022. [Online]. Available: <http://www.ijikm.org/Volume10/IJIKMv10p081-103Alarabiat0910.pdf>.
- [69] L. Leung, "Generational differences in content generation in social media: The roles of the gratifications sought and of narcissism," *Comput. Human Behav.*, vol. 29, no. 3, pp. 997–1006, 2013, doi: 10.1016/J.CHB.2012.12.028.
- [70] H. Yeatman and L. Lockyer, "Generic skills development: integrating ICT in professional preparation.," 2002, Accessed: Feb. 08, 2022. [Online]. Available: https://www.researchgate.net/publication/221093744_Generic_skills_development_integrating ICT_in_professional_preparation.
- [71] M. S. Tandale, "Use of WhatsApp as Tool for Information Dissemination in the Colleges of Western Region of Mumbai: A Study," *Int. J. Inf. Dissem. Technol.*, vol. 8, no. 3, p. 147, 2018, doi: 10.5958/2249-5576.2018.00031.6.
- [72] A. Litchfield, L. E. Dyson, E. Lawrence, and A. Zmijewska, "Directions for m-learning research to enhance active learning.," 2007.
- [73] M. Yahya Mazana, "Social Media in the classroom: WhatsApp a new communication tool for enhanced class interactions," *Bus. Educ. J.*, vol. 2, no. 1, 2018, Accessed: Feb. 08, 2022. [Online]. Available: https://www.researchgate.net/publication/332379590_Social_Media_In_The_Classroom_Whatsapp_A_New_Communication_Tool_For_Enhanced_Class_Interactions.
- [74] S. Gon and A. Rawekar, "Effectivity of E-Learning through Whatsapp as a Teaching Learning Tool," *MVP J. Med. Sci.*, vol. 4, no. 1, pp. 19–25, May 2017, doi: 10.18311/MVPJMS.V4I1.8454.
- [75] C. M. K. Cheung, P. Y. Chiu, and M. K. O. Lee, "Online social networks: Why do students use facebook?," *Comput. Human Behav.*, vol. 27, no. 4, pp. 1337–1343, Jul. 2011, doi: 10.1016/J.CHB.2010.07.028.
- [76] K. O'Hara, M. Massimi, R. Harper, S. Rubens, and J. Morris, "Everyday Dwelling with WhatsApp." Mar. 01, 2014, doi: 10.1145/2531602.2531679.
- [77] E. T. Maziriri, P. Gapa, and T. Chuchu, "Student Perceptions towards the Use of YouTube as an Educational Tool for Learning and Tutorials.," *Int. J. Instr.*, vol. 13, no. 2, pp. 119–138, Apr. 2020, doi: 10.29333/iji.2020.1329a.
- [78] J. Gulati and C. B. Williams, "Social Media in the 2010 Congressional Elections." Apr. 19, 2011, Accessed: Feb. 08, 2022. [Online]. Available: <https://papers.ssrn.com/abstract=1817053>.
- [79] A. T. Stephen, "The role of digital and social media marketing in consumer behavior," *Curr. Opin. Psychol.*, vol. 10, pp. 17–21, Aug. 2016, doi: 10.1016/J.COPSYC.2015.10.016.
- [80] B. Staehle, M. Hirth, R. Pries, F. Wamser, and D. Staehle, "YoMo: A YouTube Application Comfort Monitoring Tool," Accessed: Feb. 08, 2022. [Online]. Available: <http://www.german-lab.de/go/yomo>.
- [81] R. M. Berman et al., "The efficacy and safety of aripiprazole as adjunctive therapy in major depressive disorder: a multicenter, randomized, double-blind, placebo-controlled study," *J. Clin. Psychiatry*, vol. 68, no. 6, pp. 843–853, 2007, doi: 10.4088/JCP.V68N0604.
- [82] H. Jiang and Y. Luo, "A Dialogue with Social Media Experts: Measurement and Challenges of Social Media Use in Chinese Public Relations Practice," *Glob. Media J.*, 2012, Accessed: Feb. 08, 2022. [Online]. Available: https://www.researchgate.net/publication/283345420_A_Dialogue_with_Social_Media_Experts_Measurement_and_Challenges_of_Social_Media_Use_in_Chinese_Public_Relations_Practice.
- [83] A. A. Alalwan, N. P. Rana, Y. K. Dwivedi, and R. Algharabat, "Social media in marketing: A review and analysis of the existing literature," *Telemat. Informatics*, vol. 34, no. 7, pp. 1177–1190, Nov. 2017, doi: 10.1016/J.TELE.2017.05.008.
- [84] Y. K. Dwivedi, G. Kelly, M. Janssen, N. P. Rana, E. L. Slade, and M. Clement, "Social Media: The Good, the Bad, and the Ugly," *Inf. Syst. Front.*, vol. 20, no. 3, pp. 419–423, Jun. 2018, doi: 10.1007/S10796-018-9848-5.
- [85] D. B. Roebuck, S. M. Siha, and R. L. Bell, "Faculty Usage of Social Media and Mobile Devices: Analysis of Advantages and Concerns," *Interdiscip. J. e-Skills Lifelong Learn.*, vol. 9, pp. 171–192, 2013, doi: 10.28945/1914.
- [86] D. Shamsudeen Ibrahim, "Digitalization in Business' A Study on the Impact of Social Media Marketing Trends on Digital Marketing," *Int. J. Manag.*, 2018, doi: 10.5281/zenodo.1461321.
- [87] A. Ibrahim, D. R. Indah and D. I. Meytri, "The implementation of social customer relationship management for tourism information system," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 24, no. 3, pp. 1578-1588, 2021, doi: 10.11591/ijeecs.v24.i3.pp1578-1588.

Application of the Clahe Method Contrast Enhancement of X-Ray Images

Omarova G. S¹, Aitkozha Zh.Zh³, Nuridinov O⁶

Department of Information Systems
L. N. Gumilyov Eurasian National University
Nur-Sultan, Republic of Kazakhstan

Starovoitov V.V²

The Laboratory of System Identification
United Institute of Informatics Problems of the National
Academy of Sciences of Belarus
Republic of Belarus, Minsk

Bekbolatov S⁴

Department of Information Systems
Taraz Regional University Named after M.KH. Dulaty
Taraz, Republic of Kazakhstan

Ostayeva A.B⁵

Department of Informatics and Information and
Communication Technologies
Korkyt Ata Kyzylorda University
Kyzylorda, Republic of Kazakhstan

Abstract—Due to the nonlinearity of the luminance function produced by many medical recording devices, the quality of medical images deteriorates, which creates problems in the visual research work of physicians. X-rays can be taken as an example. This article examines methods of improving the contrast of graphic images methods of improving the quality of X-ray images. The research was carried out in several stages. Attempts were made to increase the contrast of several dozen X-ray images to select the best image brightness using brightness conversion methods in the MATLAB system. Contrast enhancement was observed during the experiments, resulting in the selection of a brightness range corresponding to the visual contrast enhancement. The selection of variables γ for the selected brightness range of the image was performed. The possibilities of the image histogram equalization method were considered. To obtain the best result before performing gamma correction the method of X-ray image histogram equalization is suggested. An enhancement version of this algorithm is presented because of the comparison. Application of the adaptive histogram equalization algorithm with contrast limitation provides a visual effect of improving the contrast of X-ray images. The NIQE and BRISQUE evaluation functions, which do not use reference images, are used to objectively quantify the conversion results.

Keywords—Digital X-ray image; image quality assessment; image enhancement; contrast enhancement; luminance transformation; adaptive image histogram equalization

I. INTRODUCTION

This One of the most powerful tools of modern informatics is medical imaging. Medical imaging is used to accurately and timely diagnose health problems, allowing patients to be treated more effectively. Nowadays, digital medical images are composed of many millions of pixels, which allows them to be considered big data. In some cases, there is a need to enhance the quality of medical images. However, in digital radiography, this may require increasing the radiation dose to the patient. Therefore, the goal of medical imaging is not to obtain a perfect image, but to obtain an image that is sufficient in terms of diagnosis concerning a particular medical problem and that causes minimal harm to the patient.

The essence of methods to enhance the quality of X-ray images is as follows: apply some mathematical methods to low contrast images and enhancement the quality of the digital medical image for a more accurate diagnosis of health problems.

II. LITERATURE REVIEW

In reviewing the experience of other researchers in this field, methods considered in the foreign literature have been studied. The paper [1] deals with contrast enhancement based on internal image decomposition, using Bregman split algorithm and CLAHE (Contrast limited adaptive histogram equalization). The authors show an enhancement meant in the images by evaluating the illumination and reflection levels using an internal image decomposition. A good contrast enhancement is obtained, but the proposed method is only for contrast enhancement and cannot be used for techniques like surface texture change, object insertion, etc.

In [2] Cheolkon Jung discusses the optimized perceptual tone mapping for contrast enhancement of images. The proposed method focuses on human visual attention by constructing a luminance histogram and performs contrast enhancement. The advantage of the method is that it enhances the performance without excessive contrast enhancement. Contrast enhancement by this method requires more time compared to HE (Histogram equalization), CLAHE methods.

S.S. Haung [3] has proposed an effective method to change the histograms and enhance the contrast of digital images. This paper presents an automatic transformation method that enhances the brightness of darkened images using gamma correction and brightness pixel probability distribution. It has been used to enhance the video data. The method proposed in the paper uses the differences between the frames to reduce the computational complexity. Experimental results have shown that the proposed method produces enhanced images of comparable or higher quality than those obtained using other methods.

M. Shakeri [4] proposed an algorithm for contrast enhancement based on local histogram equalization. The peculiarity of the algorithm is to determine the number of subhistograms and to separate the histogram based on saturation. The algorithm worked in three stages. Initially, the estimation of the number of clusters for image brightness levels is done using histogram alignment. In the next step, the image luminance levels are clustered and finally, the contrast enhancement for each cluster is included separately. The algorithm is compared with other methods based on quality and quantity measurements. The application of the method produces natural-looking images and enhanced contrast. The disadvantages of the algorithm are the loss of detail at high levels of image brightness and the presence of noise in the output image.

In work [5] the authors have proposed a new method for improving medical images. First, the original medical image is decomposed into an NSCT (contour transform without subsampling) region with a low-frequency subband and several high-frequency sub-bands. A linear transformation is then used for the luminance coefficients of the low-frequency sub-band. An adaptive thresholding method is used for noise reduction of the coefficients of the high-frequency sub-bands. All sub-bands were then reconstructed into spatial regions using the inverse NSC transform. Next, unsharp masking was applied to increase the clarity of the details of the reconstructed image. Experimental results show that the proposed method outperforms other methods in terms of such characteristics as image entropy and PSNR (peak signal to noise ratio).

In [6] paper, the social network optimized approach for image fusion for contrast enhancement and brightness preservation is discussed. The social network optimization algorithm creates two quality images, one with better contrast, and increased entropy, and the second image with an increased peak signal-to-noise ratio. The two images are combined to produce an effective image later. Comparisons were made using HE, and linear contrast stretching. The results show that the proposed method provides a better peak signal-to-noise ratio, preserves brightness, and increases the contrast of any given image, resulting in a high-quality visual effect.

However, the number of edge pixels of this technique is large, while the fit value is smaller.

In [7], Se EunKim proposes an entropy-based method for contrast enhancement in the wavelet domain. Initially, he uses local entropy scaling in the wavelet domain to obtain the desired contrast. Mathematical methods were used, and then a color enhancement method was developed in the HSI (from hue, saturation, lightness (intensity)) color space. The algorithm worked in two steps: modifying the low frequencies in the wavelet domain and scaling the HSI color space by increasing the intensity component so that images in low light get detailed color information without any further processing. The peculiarity of the algorithm is that it is used in the HSI color space and provides an increase in the contrast of the image.

Huang Lidong [8] proposed a combination of adaptive histogram equalization with limited contrast and discrete

wavelet transform to enhance the image. The algorithm works in three stages. First, the original image is allocated to low- and high-frequency components using a wavelet transform. The low-frequency coefficients are enhancements using the CLAHE method, while the high-frequency coefficients remain unchanged. When the wavelet transform is reversed, the image is mounted successfully. The proposed method is applicable for improving the local details of the image, preserving the details well, and suppressing noise. But the high-frequency component, which contains most of the noise in the original image, remains unchanged.

The authors of [9] propose a high-speed quantile-based histogram equalization (HSQHE) to preserve brightness and enhance contrast in the image. Contrast enhancement by this method is suitable for high-contrast digital images. Recursive segmentation of the histogram is not performed, so minimal time is required for segmentation. Entropy metrics are used to estimate PSNR of contrast enhancement. AMBE (Absolute Mean Brightness Error) is used to estimate brightness preservation. HSQHE preserves image brightness more accurately in a shorter time interval, but a high PSNR value is achieved only for certain images.

In [10], the authors propose a histogram modification scheme with entropy maximization. The method of histogram modification by entropy maximization divides the global histogram alignment into two stages: the pixel populations emergence (PPM) stage, which corresponds to the entropy maximization rule, and the gray-levels distribution (GLD) stage. The method gives good enhancement results and avoids reinforced noise and distortions in the image, but there is a problem with excessive contrast stretching.

The proposed methods confirm the necessity of non-linear image brightness transformation methods for contrast enhancement, but it requires detailed study for more informative images after processing.

III. IMAGE ENHANCEMENT TECHNIQUES

Image enhancement techniques involve performing such transformations on the original image that lead to a result that is more suitable for a particular application [11]. Visual assessment of image quality is an extremely subjective process, and automatic calculation of the quantitative value of such an assessment is a very difficult task. To choose one or another method to enhance the contrast of a medical image, an evaluation of the result is necessary. Objective quality assessment algorithms are divided into benchmark and non-benchmark. The different reference criteria use a comparative quality assessment when the reference image is usually known to look like, and its characteristics are known [13]. When dealing with low-contrast medical images, there are no benchmarks for comparison. Therefore, it is necessary to select those evaluation options that do not require a reference image.

Image enhancement approaches fall into two categories: spatial domain processing methods and frequency domain processing methods. The term spatial domain refers to the image plane as such, and this category combines approaches based on the direct transformation of image pixel values.

Frequency methods involve changing the images after the Fourier transform.

Let us consider some methods related to spatial processing methods. Spatial methods are described by the equation [12]:

$$g(x, y) = T[f(x, y)], \quad (1)$$

where $f(x, y)$ is a function describing the original image, $g(x, y)$ is the transformed image, and T is an operator over f defined in some neighborhood of a pixel with coordinates (x, y) . The neighborhood of a pixel is understood as a square or rectangular area that is a subset of the image and is centered relative to the given pixel. The simplest version of the T operator occurs when the neighborhood consists of a single pixel, in which case the value of g is a function of $f(x, y)$ and T is called a point type conversion.

Histogram transformations are divided into the following groups: linear logarithmic and power transformations. Histogram alignment of a digital image is a transformation of the original image in which the histogram of the transformed image has a more horizontal shape than the histogram of the original image.

To enhance image quality, it is necessary to increase such parameters as brightness range, contrast, sharpness, and sharpness. In combination, these parameters can be enhanced by aligning the histogram of the image. Histogram equalization algorithms are widely used to enhance the processed digital grayscale image. In general, such algorithms are simple to implement, have relatively low computational cost, and yet show high efficiency. The essence of such algorithms is to adjust the levels of the halftone image according to the probability distribution function of a given image (2) and, as a result, the dynamic range of brightness distribution increases. This leads to enhancement meant of visual effects, such as: brightness contrast, sharpness, and clarity.

$$P(i) = \frac{n_i}{n}, i = 0..255;$$
$$H(j) = 255 \sum_{i=0}^j P(i) \quad (2)$$

where $P(i)$ is the probability of the appearance of a pixel with brightness i , the normalized function of the histogram of the original image, j are the pixel coordinates of the processed image, $H(j)$ is the transformed image [12]. Histogram equalization algorithms are divided into the following two types: local (adaptive) histogram equalization and global histogram equalization. In the global method, one chart is built, and the histogram of the entire image is equalized. In the local method, many histograms are constructed, where each histogram corresponds to only a part of the processed image. With this method, the local contrast of the image is enhanced, which makes it possible to obtain better processing results in general.

An enhancement version of the above algorithm is the Contrast limited adaptive histogram equalization (CLAHE) algorithm. The main feature of this algorithm is the limitation of the histogram range based on the analysis of the pixel

brightness values in the processed block (3), thus the resulting image looks more natural and less noisy [14].

$$da = \frac{nc}{n} \quad (3)$$

where da is the increment factor of the value of the histogram function, nc is the number of pixels that exceed the threshold value. It is worth noting that the classic CLAHE algorithm uses bilinear interpolation to eliminate boundaries between processed blocks. The *imadjust* function is the basic tool in the MATLAB package for converting the brightness of grayscale images. All input parameters of the *imadjust* function are real numbers in the range from 0 to 1, i.e., the range of brightness values must be normalized. The syntax of the function is defined as follows:

$J = \text{imadjust}(I)$.

$J = \text{imadjust}(I, [\text{low_in}, \text{high_in}], [\text{low_out}, \text{high_out}])$.

$J = \text{imadjust}(I, [\text{low_in}, \text{high_in}], [\text{low_out}, \text{high_out}], \gamma)$ (4)

The *imadjust* function converts the intensity values of the grayscale image I to new values and writes them as a matrix J . By default, *imadjust* discards 1% of all lower and upper brightness values in the I image, then applies a linear contrast stretch.

The function $J = \text{imadjust}(I, [\text{low_in}, \text{high_in}], [\text{low_out}, \text{high_out}])$ converts the original brightness values I into new values J from the range $[\text{low_in}, \text{high_in}]$ to the range $[\text{low_out}, \text{high_out}]$. The latter can be equal to $[0, 1]$.

The function $J = \text{imadjust}(I, [\text{low_in}, \text{high_in}], [\text{low_out}, \text{high_out}], \gamma)$ additionally performs gamma correction of the converted brightness values. By default, the parameter $\gamma = 1$, which corresponds to an identical mapping [9].

Histogram equalization in MATLAB is implemented by the *histeq* function, which has the syntax:

$J = \text{histeq}(I, n)$ (5)

Where I is the input image, n is the number of intensity levels set for the output image J . If n is equal to the total number of possible levels of the input image, then *histeq* simply implements the transform function. If this number is less than the total number of possible levels of the input image, then *histeq* will redistribute the levels, so that they approximate the flat diagram. A true implementation of this method uses the maximum possible number of levels for n , which is 256. The CLAHE algorithm is implemented by the function *adapthisteq*, which has the following syntax:

$J = \text{adapthisteq}(I, \text{Name}, \text{Value})$ (6)

The Name input parameters can be:

- Number of rectangular context areas (tiles) into which *adapthisteq* divides the image, specified as a 2-element vector of positive integers;
- Contrast enhancement limit, specified as a real scalar in the range $[0, 1]$;
- Number of histogram intervals used to build a contrast-enhancing transformation (256 by default);

- Desired histogram shape;
- Distribution parameter.

CLAHE works with small areas of the image, called tiles, rather than with the whole image. The contrast of each tile is increased so that the histogram of the output area roughly corresponds to the histogram specified by the "Distribution" value. Neighboring tiles are then combined using bilinear interpolation to eliminate artificially created borders. Contrast, especially in homogeneous areas, can be limited to avoid amplifying any noise that may be present in the image.

IV. INITIAL DATA AND DESCRIPTION OF EXPERIMENTAL STUDIES

We use X-ray images from the Kaggle database [15] to experiment with the application of image brightness conversion methods. The experiment aims to increase image contrast to obtain more information about a pulmonologist's lung image representation. The essence of methods to improve the quality of medical images is to apply mathematical methods to low contrast images and improve the quality of digital medical images to improve diagnostic accuracy.

Many experiments were conducted to apply the imadjust function to several X-ray images to select the most appropriate input parameters. The values for (4) were chosen in increments of 0.1 in the range from 0 to 1 (Table 1.).

The non-referential NIQE and BRISQUE evaluation functions were used to determine how much contrast was enhanced. The NIQE (Naturalness Image Quality Evaluator) and BRISQUE (Blind/Referenceless Image Spatial Quality Evaluator) functions are used in cases where no image reference is available. The NIQE (A) function compares the quality of image A relative to an abstract model image constructed from images of natural scenes. The BRISQUE (A) function compares the quality of image A with respect to another model image constructed from many images of natural scenes with certain distortions. The smaller the values of these functions, the higher the quality of the images.

TABLE I. SELECTION OF PARAMETER VALUES OF THE IMADJUST FUNCTION

Image title	Selected imadjust options	Original Score		Post-conversion assessment	
		Niqe	Brisque	Niqe	Brisque
1.png	[0.4,1] [0,1]	4.0372	16.1975	3.4770	32.7370
2.png	[0.5,1] [0,1]	4.2881	18.7059	3.8257	32.7584
3.png	[0.2,1] [0,1]	4.1413	10.4101	3.9845	32.8306
4.png	[0.3,1] [0,1]	4.2956	13.0724	3.8182	32.3951
5.png	[0.2,1] [0,1]	4.3203	25.7744	3.8746	33.5517
Normal	[0.1,1] [0,1]	3.1248	18.1867	2.7623	25.4380
Pneumonia1	[0.3,1] [0,1]	3.0242	36.0416	2.6395	36.6267
Pneumonia2	[0, 1][0, 1]	2.7003	34.2984	2.7003	34.2984
Pneumonia3	[0.2,1] [0,1]	3.0398	13.8546	2.9204	33.3662
Pneumonia4	[0.2,1] [0,1]	3.0501	45.8458	2.9693	42.2854

During the experiments we have tried many brightness ranges of source images, for which the attempts to increase the contrast of the X-ray images gave a positive result both visually and quantitatively. Table 1 shows examples of imadjust function parameters in determining the most appropriate value of parameter γ . If $\gamma < 1$, the resulting image will be lighter than the original one. However, in most cases there was no positive result in improving the image. If $\gamma > 1$, the curve of transformation of brightness values will be concave, and the resulting image will be darker than the original one. Here, for each selected value [low_in, high_in], [low_out, high_out], the parameter γ was selected from the range [1, 44.5] in increments of 0.5. From all [low_in, high_in] [low_out, high_out] the ones with the best γ values were selected, then they were compared in the table. For example, for image 1.png the results are shown in Table 2.

Thus, according to the data in Table 2, you can determine the best value of the input parameters of the imadjust function. When you select these parameters, you can visually display the result of the transformation and compare it with the original image (Fig. 1).

Figure 1. shows the original image(a) and the result of applying the imadjust function with the selected parameters(b). Here the NIQE score for the original image is 4.0372 and for the transformed image the score is 3.3252. We can note the higher contrast of the transformed image and the NIQE score shows a lower value than that of the original image.

Table 3 shows the best score values for the 10 test images.

When selecting the value of parameter γ , in most cases of performing the function, the result of transformation did not give improvement and visual perception and in the quantitative assessment of the result. For example, Figure 2 shows the results of the transformation of the original image 4.png.

TABLE II. RESULTS OF APPLYING THE PARAMETER Γ

Imadjust input parameters	Estimates for $\gamma = 1$		Best gamma value	Estimation	
	Niqe	Brisque		Niqe	Brisque
[0.2 1] [0 1]	3,8665	13,0019	$\gamma=4$	3,3524	16,0788
[0.3 1] [0 1]	3,7775	14,7344	$\gamma=2,5$	3,3252	26,8559
[0.4 1] [0 1]	3,4770	32,7370	$\gamma=2,5$	3,3790	25,9206
[0.5 1] [0 1]	3,6238	31,6125	$\gamma=2,5$	3,5020	28,2055

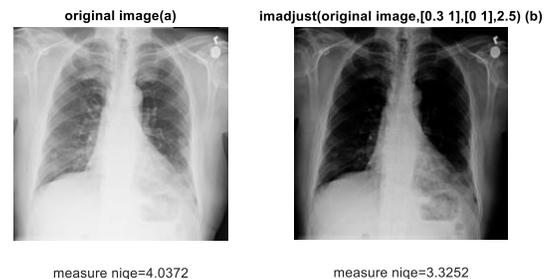


Fig. 1. Comparison of Imadjust ('1.png',[0.3 1],[0 1],2.5) (b) with the Original Image (a).

TABLE III. CHOICE OF Γ PARAMETER VALUE

Image title	brightness options	Estimates for $\gamma=1$		The best value of the parameter γ	Niqa evaluation	Brisque Evaluation
		Niqa	Brisque			
1.png	[0.3 1] [0 1]	3,7775	14,7344	$\gamma=2.5$	3,3252	26,8559
2.png	[0.4 1] [0 1]	3.9106	32.5165	$\gamma=2$	3.6851	22.3536
3.png	[0.2 1] [0 1]	3.9845	32.8306	$\gamma=2$	3.8189	25.2399
4.png	[0.2 1] [0 1]	4.1986	36.9663	$\gamma=2$	3.8848	25.4878
5.png	[0.2 1] [0 1]	4.0250	37.2250	$\gamma=2$	3.8306	31.3175
N	[0.2 1] [0 1]	3.2911	33.5240	$\gamma=2$	3.3236	21.5805
P	[0.3 1] [0 1]	2.6395	36.6267	$\gamma=1.5$	2.6767	37.6345
P	[0 1] [0 1]	3.2048	41.9046	$\gamma=2.5$	3.2404	41.8697
P	[0.2 1] [0 1]	2.9204	33.3662	$\gamma=2$	2.5273	21.3953
P	[0.2 1] [0 1]	2.9693	42.2854	$\gamma=2$	3.0508	39.3603

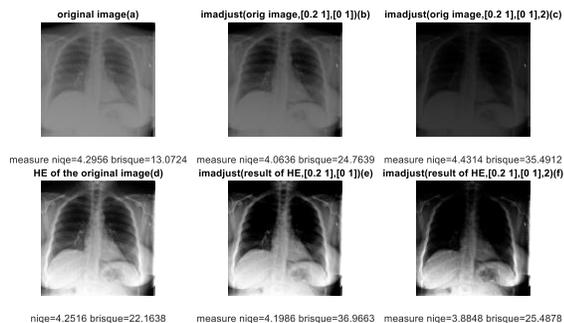


Fig. 2. Original Image (a) and its Transformed Versions with Estimates (b)(c)(d)(e)(f).

Applying histogram equalization (5) of the original image before testing the imadjust function with the choice of the parameter γ , gives the result of enhancement image contrast (Table 4).

In the following experiment, histogram equalization techniques are applied to several images, comparing their results with the quality of the original image. For example, for the above image 4.png (a), the application of histogram equalization (b) and adaptive histogram equalization with contrast restriction (c) are shown in Figure 3.

Figure 4 shows the results of similar actions for another image person9_bacteria_39.jpeg. It can be seen in the figures that the application of the adaptive histogram equalization method with contrast restriction (c) compared to the HE images result (b) visually gives a better result, but the NIQE and BRISQUE estimates do not always match.

TABLE IV. IMAGE ESTIMATES AFTER HISTOGRAM EQUALIZATION

Brightness conversion	Estimates	
	Niqa	Brisque
source image(4.png)	4.2956	13.0724
imadjust(original,[0.2 1],[0 1])	4.0636	24.7639
imadjust(source,[0.2 1],[0 1],2)	4.4314	35.4912
Alignment of the histogram of the original image	4.2516	22.1638
imadjust(original aligned,[0.2 1],[0 1])	4.1986	36.9663
imadjust(source aligned,[0.2 1],[0 1],2)	3.8848	25.4878

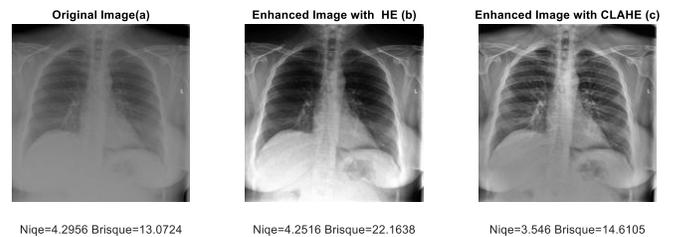


Fig. 3. Comparison of the Results of Applying Histogram Equalization Methods to the Image with Non-reference Estimates for Image 4.png.

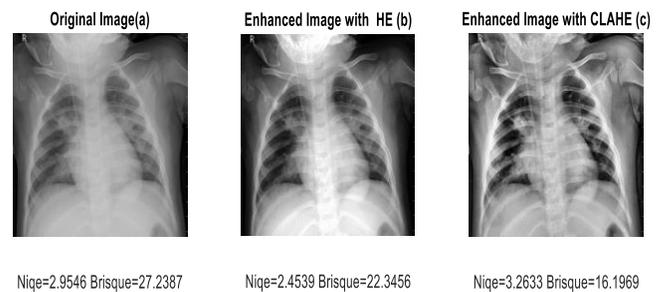


Fig. 4. Comparison of the Results of Applying Image Histogram Equalization Methods with Non-reference Estimates for the Person9_bacteria_39.jpeg (a) Image.

Table 5 shows the scores of 15 test images after applying the histogram equalization methods. In most cases, the results of applying the CLAHE method show a visual improvement in image contrast and a reduction in non-reference scores at the same time. In some cases, the estimates of the results of applying contrast-limited adaptive equalization do not decrease in value compared to the estimates of the original image.

As a result of analyzing the data in Table 5, it was decided that to improve the results of image contrast enhancement, it would be appropriate to replace the histogram equalization method with adaptive histogram equalization with contrast restriction. In the following experiment, function (6) was used to improve the contrast of image I in grayscale by transforming the values using adaptive histogram equalization with contrast restriction.

The application of this method was focused on the Distribution and SlipLimit parameters. The Distribution parameter takes the values 'uniform', 'rayleigh', 'exponential', which set the desired shape of the histogram. This parameter defines the distribution that adapthisteq uses as the basis for

creating the contrast conversion function. The selected distribution should depend on the type of input image. For example, underwater images seem more natural when using the 'rayleigh' distribution.

TABLE V. IMAGE SCORES AFTER APPLYING HISTOGRAM EQUALIZATION METHODS

Image title	Original image		Histogram equalization result		CLAHE result	
	Niqe	Brisque	Niqe	Brisque	Niqe	Brisque
1.png	4.03 72	16.19 75	3.80 41	18.59 71	3.27 15	10.64 72
2.png	4.28 81	18.70 59	4.07 96	25.81 75	3.38 52	6.668 7
3.png	4.14 13	10.41 01	4.84 12	29.74 37	3.40 34	8.295 1
4.png	4.29 56	13.07 24	4.25 16	22.16 38	3.54 60	14.61 05
5.png	4.32 03	25.77 44	3.85 08	27.60 71	3.85 08	27.60 71
6.png	4.80 23	29.95 13	5.40 88	40.31 79	4.22 07	28.35 85
person1_bacteria_2.jpeg	3.08 89	28.76 98	2.52 52	26.22 16	3.37 20	12.78 19
person2_bacteria_4.jpeg	3.34 58	19.78 43	3.06 30	20.61 80	3.88 28	24.87 27
person3_bacteria_10.jpeg	2.83 16	21.72 51	2.91 40	22.74 25	3.15 78	21.87 98
person5_bacteria_15.jpeg	2.43 08	34.78 98	2.34 27	32.89 20	2.95 93	28.56 70
person6_bacteria_22.jpeg	2.63 89	29.06 88	2.38 90	19.89 88	3.32 71	17.62 35
person7_bacteria_24.jpeg	2.81 25	28.86 72	2.56 47	28.13 11	3.09 36	2.235 6
person8_bacteria_37.jpeg	2.76 26	31.06 23	2.33 59	29.35 76	2.33 59	29.35 76
person9_bacteria_39.jpeg	2.95 46	27.23 87	2.45 39	22.34 56	3.26 33	16.19 69
person17_bacteria_56.jpeg	2.69 56	38.59 77	2.69 56	38.59 77	2.69 56	38.59 77

The ClipLimit parameter is a contrast ratio that prevents oversaturation of the image, especially in homogeneous areas. These areas are characterized by a high peak on the histogram of a particular image fragment because many pixels fall within the same range of gray levels. Without clip limitation, the Adaptive Histogram Smoothing method can produce results that are, in some cases, worse than the original image. Its default value is 0.01.

The following steps were performed for several test X-ray images:

- To determine the optimal value of the 'clipLimit' parameter, we chose its values from the interval [0, 1] in steps of 0.01.

- Calculation of objective estimates for all transformed images.
- Plotting objective estimates for all versions of images.
- Determination of the minimal estimates NIQE and BRISQUE;
- Choosing the optimal visual representation of the image with the minimum objective estimates.

Construction of objective scores plots (Fig. 5) for several X-ray images showed that the values of cliplimit parameter can be limited from [0, 1] to [0, 0.2], as the following values were not informative. The minimal measures of the NIQE and BRISQUE estimates allow us to select images with improved contrast. This choice is related to the claim that the smaller the value of the non-reference score, the visually improved the image is. This assertion has been proven in previous studies, where a minimum NIQE score was more likely to coincide with an improved visual perception of the image.

Figure 6 shows a visual comparison of the original image (a) with the transformed one (b), where the clahe method is applied with the selected parameters and with the minimal NIQE score. Here the value of the distribution parameter is equal to 'rayleigh' and those obtained images are selected, at which the non-reference estimates had minimal values. For example, for the image 1.png the minimal estimate NIQE=2.9012 was received at cliplimit=0.12, and the BRISQUE=15.314 corresponds to it. For the image with minimal BRISQUE score equal to 9.1993 at value of cliplimit=0.01 the NIQE=3.2265 was defined. Here it can be noted that the decrease of the BRISQUE score in many cases does not correspond to the decrease of the NIQE score at which visual improvements were observed.

A visual comparison of the original image (a) with the CLAHE-transformed image (b) with minimal BRISQUE estimation is shown in Figure 7. Here the parameter distribution at value 'rayleigh' takes minimum BRISQUE value equal to 9.1993, which corresponds to NIQE=3.2365 at value of parameter cliplimit=0.01.

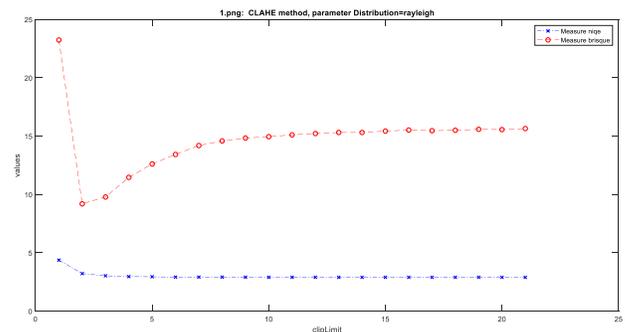


Fig. 5. Plots of Objective Estimates for the Transformed Images of the Original '1.png' with Distribution='Rayleigh'; and 'ClipLimit'=[0,0.2] with Step 0.01 (BRISQUE Estimates Marked in Red, Niqe Estimates Marked in Blue).

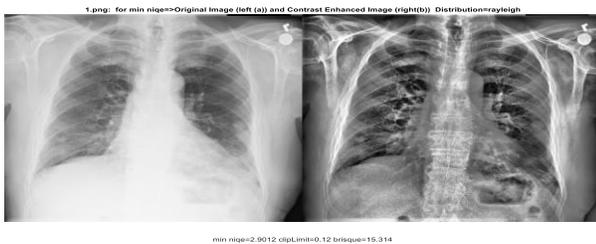


Fig. 6. Comparison of the Result of the Transformation of the Original Image (a) by the CLAHE Method (Distribution='Rayleigh', ClipLimit=0.12) (b) with the Minimum NIQE Estimate.

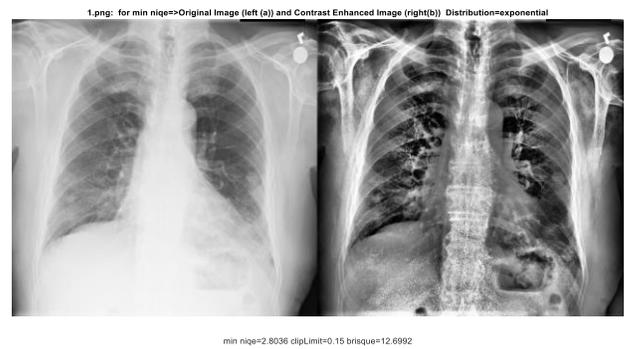


Fig. 9. Visual Comparison of the Original Image (a) with the Transformed CLAHE Method (Distribution='Exponential', ClipLimit=0.15)(b) and with the Minimum NIQE Estimate.

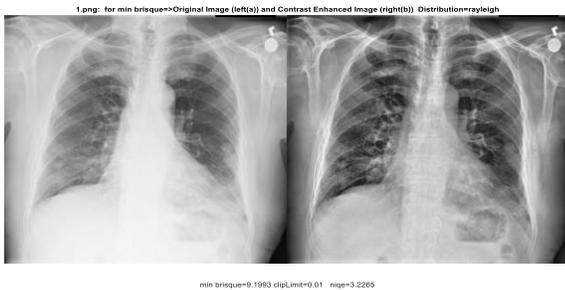


Fig. 7. Visual Comparison of the Original Image (a) with the Transformed CLAHE Method (Distribution='Rayleigh', ClipLimit=0.01) (b) and with the Minimum BRISQUE Estimate.

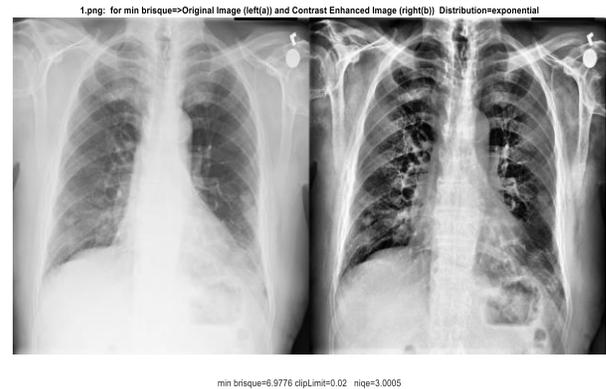


Fig. 10. Visual Comparison of the Original Image (a) with the CLAHE Transformed Image (Distribution='Exponential', ClipLimit=0.02) (b) with Minimum BRISQUE Estimation.

The plots of the objective estimates for the transformed images of the original '1.png' by the adaptive histogram equalization method with contrast constraint are shown in Figure 8. Here the distribution parameter takes the value 'exponential'; and the parameter 'clipLimit' receives values from the interval [0,02] with a step of 0.01.

A visual comparison of the original image (a) with the CLAHE-transformed image (b) with the minimum BRISQUE score is shown in Figure 9. Here the parameter distribution with 'exponential' value takes a minimum NIQE value of 2.8036, which corresponds to BRISQUE=12.6992 with the value of the parameter clipLimit=0.15.

A visual comparison of the original image (a) with the CLAHE-transformed image (b) with minimal BRISQUE estimation is shown in Figure 10. Here the parameter distribution at value 'exponential' takes minimum BRISQUE value equal to 6.9796, to which corresponds NIQE=3.0005 at value of parameter clipLimit=0.02.

The results of similar actions performed on the rest of the test images are shown in Table 6. Here are the non-reference estimates of the original image and the CLAHE transformation results with the selected values of the distribution parameter. For each value of this parameter, the minimum estimates of NIQE and BRISQUE, and their corresponding values of the clipLimit parameter and estimates have been determined.

Table 6 shows the values of the obtained non-referential estimates of the original image and the transformed images using the CLAHE method. Changing the values of the distribution and clipLimit parameters, when performing the adaptive equalization method with contrast restriction, gives positive results. An analysis of the values in Table 6 gives a preference for the value of the 'distribution='exponential' parameter for certain values of the clipLimit parameter. This is evidenced by the NIQE and BRISQUE non-reference scores, which decrease in value as the contrast of medical images improves. As demonstrated by the laboratory studies performed, in many cases the NIQE score was more consistent with image improvement.

As a result of the performed laboratory studies, a combination of the gamma correction method and the adaptive histogram equalization method, in which contrast enhancement is limited to avoid causing or enhancing noise in the image, is considered appropriate.

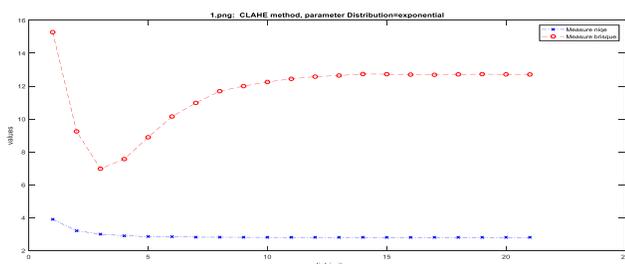


Fig. 8. Plots of Objective Estimates for the Transformed Images of the Original '1.png' with Values of Distribution='Exponential'; and 'ClipLimit'=[0,02] in Steps of 0.01 (BRISQUE Estimates Marked in Red, NIQE Estimates Marked in Blue).

TABLE VI. COMPARISON OF NON-REFERENCE ESTIMATES OF THE ORIGINAL IMAGE AND THE CLAHE-TRANSFORMED IMAGES WHEN CHANGING THE VALUES OF THE DISTRIBUTION AND CLIPLIMIT PARAMETERS

Image	Niqa (original)	Brisque (original)	distribution	min Niqa	for min Niqa, cliplimit	for min Niqa, Brisque	min Brisque	for min Brisque, cliplimit	for min Brisque, Niqa
1	4.0372	16.1975	'rayleigh'	2.9012	0.1200	15.314	9.1993	0.0100	3.2265
			'exponential'	2.8036	0.1500	12.6992	6.9776	0.0200	3.0005
2	4.2881	18.7059	'rayleigh'	3.0420	0.0800	15.7290	8.9939	0.0100	3.3514
			'exponential'	3.0024	0.0800	14.7401	7.2666	0.0100	3.3447
3	4.1413	10.4101	'rayleigh'	3.1609	0.0700	14.4351	6.6493	0.0100	3.4322
			'exponential'	3.0930	0.0700	15.6488	9.0976	0.0100	3.3438
4	4.2956	13.0724	'rayleigh'	3.2971	0.1700	17.8653	13.0724	0.0100	3.5975
			'exponential'	3.2193	0.1700	19.9392	13.0724	0.0100	3.5217
5	4.3203	25.7744	'rayleigh'	2.9495	0.0500	27.6091	25.7744	0.0100	3.3356
			'exponential'	2.9055	0.0600	26.7410	22.3760	0	4.2776
6	4.8023	29.9513	'rayleigh'	3.9037	0.1300	17.1803	16.9361	0.2300	3.9085
			'exponential'	3.9655	0.1600	19.0927	18.9781	0.2100	3.9714
7	3.0889	28.7698	'rayleigh'	3.0759	0	33.8095	4.9285	0.0100	3.1913
			'exponential'	3.0337	0	29.4685	10.0346	0.0100	3.3622
8	3.3458	19.7843	'rayleigh'	3.2490	0	19.1865	19.1865	0	3.2490
			'exponential'	3.3083	0	13.8267	13.8267	0	3.3083
9	2.8316	21.7251	'rayleigh'	2.6969	0	28.6374	11.0289	0.0100	3.0216
			'exponential'	2.7980	0	21.5236	18.7296	0.0100	3.1852

V. CONCLUSION

During the experiment, X-ray images were used, some of which visually improved without difficulty during luminance conversion, some of which took a darker shade after conversion, and the image quality remained poor. When working with such images, it was difficult to improve the contrast using gamma correction. To achieve better contrast, an image histogram alignment was performed before applying gamma correction. This resulted in better results. Based on the final Table 3, we can conclude that the best results were achieved with the input parameters [0.2 1] [0 1] with $\gamma = 2$. As a result of research of test image transformation variants, to improve the contrast of X-ray images it is recommended first to apply the histogram equalization procedure and then imadjust transformation with the parameters ([low_in 1] [0. 1], 2), where $0.2 \leq \text{low_in} \leq 0.4$. To improve the obtained results, it was decided to replace the histogram equalization with adaptive histogram equalization with contrast limitation. Because of applying this method, it was determined that the 'exponential' value is given preference when the distribution parameter is given a value values of the cliplimit parameter. It was also determined during the research that in most cases the quantitative measure of NIQA is more consistent with image improvement than the BRISQUE score when evaluating image quality.

REFERENCES

- [1] Huanjing Yue, Jingyu Yang, Xiaoyan Sun, Feng Wu. Contrast Enhancement Based on Intrinsic Image Decomposition, IEEE Transactions on image processing 2017, 26(8), P.3981-3994.
- [2] Cheolkon Jung, Tingting Sun. Optimized Perceptual Tone Mapping for Contrast Enhancement of Images, IEEE Transactions on circuits and systems for video technology 2017, 27(6), P. 1161-1170.
- [3] S.S. Haung, F.S. Cheng, Y.C. Chiu. Efficient contrast enhancement Using Adaptive Gama Correction with Weighting Distribution. IEEE Transactions on Image Processing 2013; 22 (3): P.1032-1041.
- [4] M.Shakeri, M.H.Dezfoulian, H.Khotanlou, A.H.Barati, Y.Masoumi. Image contrast enhancement using fuzzy clustering with adaptive cluster parameter and sub-histogram equalization", Elsevier Digital signal Processing 2017, P. 224-237.
- [5] L.Liu, Z. Jia, J. Yang, N. Kasabov. A Medical Image Enhancement Method Using Adaptive Thresholding in NSCT Domain Combined Unsharp Masking. Wiley Periodicals, Inc. 2015, 25: P.199-205.
- [6] Lalit Maurya, Prasant Kumar Mahapatra, Amod Kumar. A social spider optimized image fusion approach for contrast enhancement and brightness preservation, Elsevier Applied soft computing 2017, P.575-592.
- [7] Se EunKim, JongJuJeon, IlKyuEom. Image contrast enhancement using entropy scaling in wavelet domain, Elsevier signal Processing 2016, P. 1-11.
- [8] Huang Lidong, Zhao Wei, Wang Jun, Sun Zebin, Combination of contrast limited adaptive histogram equalisation and discrete wavelet transform for image enhancement, IET Image Processing Journals 2015, Vol. 9, Iss. 10, P. 908-915.
- [9] Mayank Tiwari, Bhupendra Gupta, Manish Shrivastava, Highspeed quantile-based histogram equalisation for brightness preservation and contrast enhancement, IET Image Process 2015, 9(1), P. 80-89.

- [10] Zhao Wei, Huang Lidong, Wang Jun, Sun Zebin, Entropy maximisation histogram modification scheme for image enhancement, *IET Image Process* 2015, 9(3), P. 226–235.
- [11] Gonzalez R., Woods R. *Digital image processing. - 3rd edition, revised and supplemented.* –Moscow: Technosphere, 2012. – 1104 p.
- [12] Gonzalez R., Woods R., Eddins S. *Digital image processing in Matlab.* – M.: Technosphere, 2006.-616 p.
- [13] Starovoitov F.V., Parameters of the distribution curve of local estimates as a measure of image quality / F. V. Starovoitov, V. V. Starovoitov // *System analysis and applied informatics.* – 2018. – No. 3. – pp. 26-41.
- [14] Ma J., Fan X., Yang S.X., Zhang X., Zhu X. Contrast Limited Adaptive Histogram Equalization Based Fusion for Underwater Image Enhancement // *Preprints [Электронный ресурс]* 2017, URL: <https://www.preprints.org/manuscript/201703.0086/v1>.
- [15] <https://www.kaggle.com/paultimothymooney/chest-xray-pneumonia>.

Efficient Segment-based Image Ciphering using Discretized Chaotic Standard Map with ECB, OFB and CBC

Mohammed A. AlZain

Department of Information Technology, College of Computers and Information Technology
Taif University, P.O. Box 11099, Taif 21944, Saudi Arabia

Abstract—This paper presents a block-based ciphering scheme that employs the 2D discretized chaotic Standard map (CSM) in three different operation modes. The employed operation modes include the electronic codebook (ECB), the output feedback (OFB) and the cipher block chaining (CBC) modes. In the presented 2D discretized CSM with the OFB and CBC, the initiation vector (IV) is employed as the primary secret key. The presented 2D discretized CSM with the CBC has two merits. The first merit is the ability of the presented 2D discretized CSM with the ECB, OFB and CBC to encipher images of any dimensions in a comparatively short time. The second merit is the high level of security of the presented 2D discretized CSM with the OFB and CBC through the integration of both diffusion and confusion operations. Different security key metrics like histogram deviation, irregular, and coefficient of correlation, are examined to assess the functionality of the presented 2D discretized CSM with the OFB and CBC. The resistance to noise, uniformity of histogram, and encryption speed are also investigated. The suggested 2D discretized CSM with the OFB and CBC is compared with the 2D discretized CSM in ECB. The achieved outcomes demonstrate that the proposed 2D discretized CSM with the OFB and CBC has a high security than in ECB from cryptographic viewpoint. Also, achieved outcomes demonstrate that the proposed 2D discretized CSM has better noise immunity in OFB compared with ECB and OFB.

Keywords—Cryptography; 2D discretized CSM; ECB; OFB; CBC

I. INTRODUCTION

The cryptography field is especially important in the modern era, where information security is paramount. Security is an important issue for image communications and storage, and ciphering is considered as one of the most important ways to realize and ensure security. Ciphering has a lot of applications such as online communications, multimedia communications, medical image protection, telemedicine, military communications, pay TV, and video conferencing. Chaotic ciphering has an important role in current cryptography. The attraction of utilizing chaotic ciphering for implementing the current cryptosystems is due to several reasons which satisfy the traditional Shannon requirements of both diffusion and confusion [1-3]. These reasons may include its random-like behavior and parameters setting sensitivity and preconceived conditions [4]. Chaotic-based schemes have demonstrated some positive features in many of the affected

areas in terms of speed, integration complexity, security, power, and computational overhead. Now, some ciphers for securing images have been presented [5-8]. Other cryptosystems based on discrete chaotic systems have been suggested, but still they have some concerns about their security [9-15].

Actually, there exists two basic ways to approach digital images chaotic ciphering. In the first approach, a chaos-based stream cipher model is employed for generating pseudo-random access keys to hide the source plaintext [16]. This model is known as stream cipher. In the other scheme, the source text or secret key can be employed as initialization conditions or control parameters, and the chaotic system is iterated for several rounds to deliver the final encrypted data [17-18]. This model is known as block cipher. Now, the chaotic 2D maps have been developed into 3D to design symmetric cryptosystems, which are intended to increase security. The 3D chaotic Baker mapping introduced by Mao et al. and 3D chaotic Cat mapping introduced by Chen et al. [19-20] represent some examples of 3D chaotic maps.

In the article, a block-based ciphering scheme that employs the 2D discretized CSM with the ECB, OFB and CBC is presented. In the presented 2D discretized CSM with the OFB and CBC, the image is split into blocks, and the encryption is applied for each one of these blocks using the 2D discretized CSM with the ECB, CBC and OFB. The initiation vector (IV) is employed as the primary secret key. The 2D discretized CSM with the CBC provides two advantages. The first is the ability of the presented 2D discretized CSM with the ECB, OFB and CBC to encipher images of any dimensions in a comparatively short time. The second one is the high level of security of the presented 2D discretized CSM with the ECB, OFB and CBC through the integration of both diffusion and confusion operations. Several security key performance metrics like histogram deviation, irregular, and coefficient of correlation, are examined to assess the functionality of the presented 2D discretized CSM with the ECB, OFB and CBC. The resistance to noise, uniformity of histogram, and encryption speed are also investigated. The suggested 2D discretized CSM with the ECB, OFB and CBC is compared with the 2D discretized CSM and the RC5. The achieved outcomes demonstrate that the proposed 2D discretized CSM with the ECB, OFB and CBC has a high security from cryptographic viewpoint.

The remainder of this paper is structured as follows. Section II provides an overview of the traditional 2D CSM, 2D discretized CSM in addition to the utilized ECB, OFB and CBC operation modes. Section III presents the introduced image cryptosystem using 2D discretized CSM with the ECB, OFB and CBC. Section IV gives the design issues of the proposed image cryptosystem using 2D discretized CSM with the ECB, OFB and CBC cryptosystem. Section V provides encryption quality metrics used to evaluate the performance of the proposed image cryptosystem using 2D discretized CSM with the ECB, OFB and CBC. Section VI provides the results of the presented image cryptosystem using 2D discretized CSM with the ECB, OFB and CBC. Finally, the paper conclusions are listed in Section VII.

II. PRELIMINARY TOOLS

This part provides a compact overview for three conventional cryptosystems, the 2D CSM, 2D discretized CSM, and RC5 ciphers in addition to the utilized ECB, OFB and CBC operation modes. All of 2D CSM, 2D with the ECB, OFB and CBC are symmetric block ciphers. In both crypto ciphers, the utilized key is the same for both of encryption and decryption.

A. The 2D Chaotic Standard Map (2D CSM) and 2D Discretized CSM Cipher

With chaos-based image ciphering, the positions of pixels are arbitrarily changed. Different chaotic-based maps may be employed with chaos-based image ciphering like 2D Cat, 2D Henon, 2D Baker, line, and General maps. The Standard mapping, Cat mapping, Henon mapping, and Baker mapping employ processes of geometric modifications. The line mapping employs stretching of the whole pixels form a straight line, and employs folding according to certain rules. Then, the plainimage pixels are chaotically distributed in the resulted cipherimage and nearby pixels are no longer important. On contrary, a typical 2D CSM was developed by Boris Chirikov in 1969. The 2D CSM with continuous confusion is described as follows [21]:

$$\begin{bmatrix} u(i+1) \\ v(i+1) \end{bmatrix} = \begin{bmatrix} (u(i) + v(i)) \bmod 2\pi \\ (v(i) + k \sin u(i+1)) \bmod 2\pi \end{bmatrix} \quad (1)$$

If Eq. 1 is discretized, it will be mapped from $[0, 2\pi]$, to $M \times M$ through putting $u = uM/2\pi$, $v = vM/2\pi$, and $k = kM/2\pi$ to transform the 2D CSM to the 2D discretized CSM as follows [21]:

$$\begin{bmatrix} u(i+1) \\ v(i+1) \end{bmatrix} = \begin{bmatrix} (u(i) + v(i)) \bmod M \\ \left(v(i) + K \sin \frac{u(i+1)M}{2\pi} \right) \bmod M \end{bmatrix} \quad (2)$$

where k denotes a non-negative integer.

If the 2D discretized CSM is employed for image ciphering, $u(i)$ and $v(i)$ represent the pixel coordinates of the plainimage. $u(i+1)$ and $v(i+1)$ represent the pixel coordinates of the cipherimage.

The 2D discretized CSM is intended to achieve continuous map properties; it must be very close to the base map as the number of pixels is usually endless [21]. The resulted cipher of the 2D discretized CSM is a permutation cipher, which cannot modify the cipherimage histogram from its corresponding plainimage. Since this cipher is simple and fast, it does provide a high level of security, and its processing time grows as image dimensions increases.

B. The ECB Mode

The ECB mode starts through segmenting the input data into segments of equal sizes as illustrated in Fig. 1(a). Then every segment is separately encrypted using the same encryption key. The ECB operation mode can be mathematically represented using the following equation:

$$\text{Cipher}_j = \text{ENC}_k(\text{Plain}_j), \quad j=1,2,3,\dots,n \quad (3)$$

The ECB deciphering process can be represented as:

$$\text{Plain}_j = \text{DEC}_k(\text{Cipher}_j), \quad j=1,2,3,\dots,n \quad (4)$$

C. The OFB Mode

The OFB mode starts through ciphering the IV as illustrated in Fig. 1(b). Then, the resulted output bits are XORed with their corresponding plaintext block to result in the ciphertext block. In addition, the resulted ciphertext block bits are utilized an input IV to the next stage. The procedure is repeated till reaching the final block. The OFB operation mode can be mathematically represented using the following equation:

$$\text{Cipher}_j = \text{Plain}_j \oplus I_j, \quad j=1,2,3,\dots,n \quad (5)$$

The OFB deciphering process can be represented as:

$$\text{Plain}_j = \text{Cipher}_j \oplus I_j, \quad j=1,2,3,\dots,n \quad (6)$$

where $I_j = \text{ENC}_k(I_{j-1})$, $j=1, 2, 3 \dots n$, and $I_0 = \text{IV}$.

As CBC mode, ciphering phase should be strong enough to provide efficient immunity to any attack attempts to break it.

D. The CBC Operation Mode

The CBC operation mode is a segment encryption mechanism as illustrated in Fig. 1(c). The CBC operation mode has been employed for use with the 2D discretized CSM in the introduced cipher. In the introduced 2D discretized CSM cipher with CBC, the CBC mode utilizes IV of equivalent size to the segmented block size. First, each one of IV pixels is XORed with its corresponding block pixel in the 1st block. After that, the outgoing pixels are ciphered. The first block pixels are employed as IV to encipher the second block. The process is repeated with the same sequence until reaching the final block. The CBC operation mode can be mathematically represented using the following equation:

$$\text{Cipher}_j = \text{ENC}_k(\text{Cipher}_{j-1} \oplus \text{Plain}_j), \quad j=1, 2, 3, \dots, n \quad (7)$$

where $\text{Cipher}_0 = \text{IV}$, Cipher_j denotes the ciphered block, \oplus represents the XOR operation, and ENC_k denotes the 2D CSM encryption.

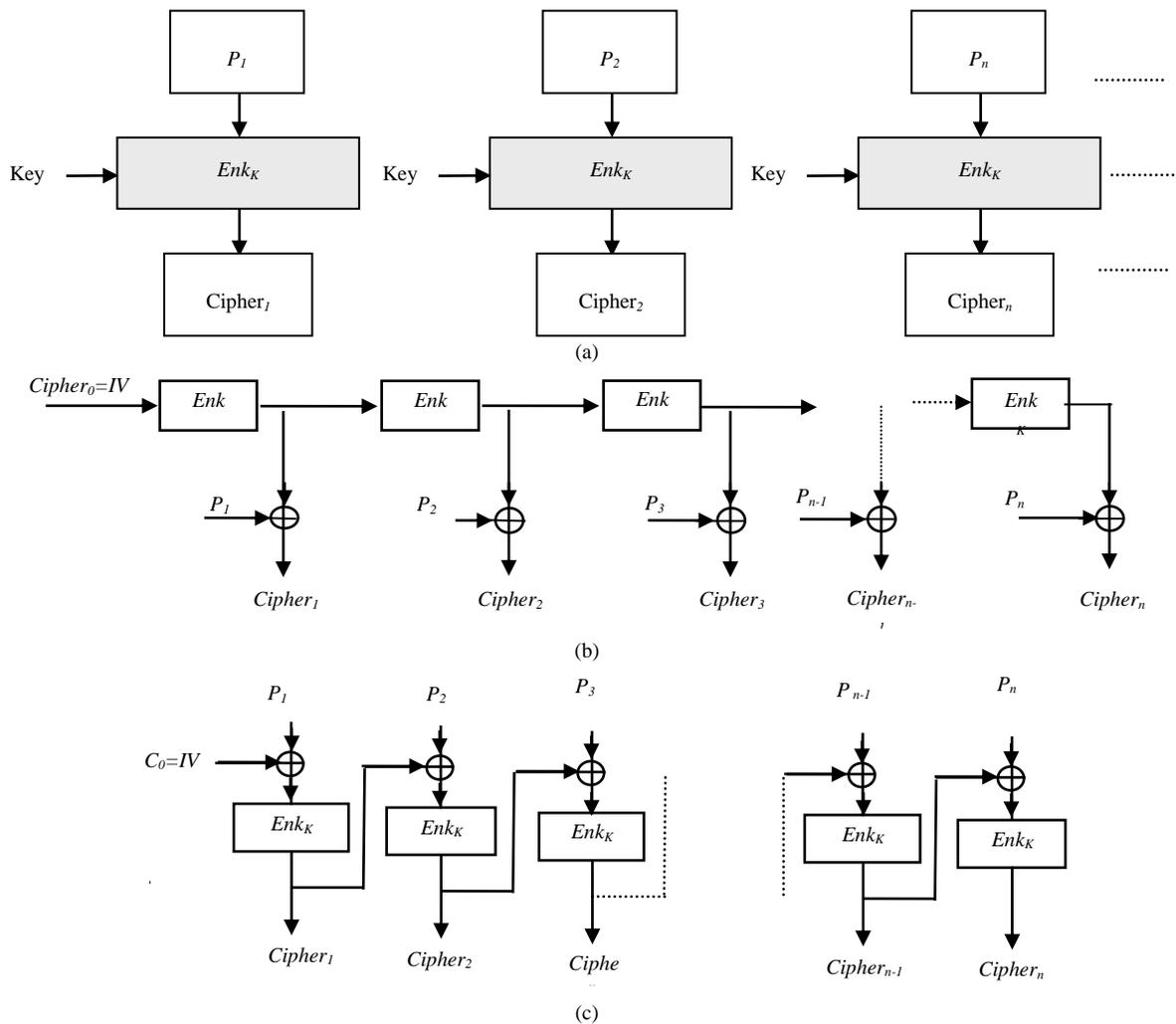


Fig. 1. The Employed ECB, OFB and CBC with Introduced 2D Discretized CSM Cipher as Depicted in (a), (b) and (c).

The CBC operation mode employs an integration method that employs dependence on each cipherimage block for all the previous cipherimage blocks. As a consequence, all eligibility for all previous cipherimage blocks is contained in the previous cipherimage blocks [22]. The CBC basic drawback disadvantage lies in the fact that an attack on just single cipherimage segment affects two plainimage segments when employing decryption [23]. The CBC deciphering process can be represented as:

$$\text{Plain}_j = \text{DEC}_k(\text{Cipher}_j) \oplus \text{Cipher}_{j-1}, \quad j=1, 2, 3, \dots, n \quad (8)$$

where DEC_k denotes the deciphering procedure.

III. PROPOSED 2D DISCRETIZED CSM-BASED IMAGE CIPHER WITH OFB AND CBC OPERATION MODES

This part is to provide an overview of the introduced 2D discretized CSM cipher with the ECB, OFB and CBC. The proposed 2D discretized CSM cipher with the ECB, OFB and CBC is designed with the potential of enhancing the security the cipherimage and providing a reasonable encryption/decryption speeds. For realizing these objectives, 2D discretized CSM encryption is employed with the ECB, OFB and CBC [24-29]. Three schemes of the 2D discretized

CSM with ECB, OFB and CBC are examined to determine which operation mode that will increase the efficiency of the proposed 2D discretized CSM cipher.

The operation of the proposed 2D discretized CSM cipher with the ECB, OFB and CBC may be summed up as shown below in the next three steps. The steps of the proposed 2D discretized CSM cipher with the ECB, OFB and CBC is depicted can be listed as:

- 1) The plainimage to be encrypted is scanned line by line.
- 2) The scanned image is segmented into blocks of; each block of has $n \times n$ pixels.
- 3) The segmented image blocks are ciphered using the proposed 2D discretized CSM cipher with the ECB, OFB and CBC modes of operation as depicted in Fig. 1.

IV. DESIGN POINTS OF THE 2D DISCRETIZED CSM CIPHER WITH THE OFB AND CBC

As mentioned previously, the proposed 2D discretized CSM cipher may be employed in ECB, OFB and CBC operating modes. It basically employs a 2D discretized CSM with the ECB, OFB and CBC as the main cipher scheme. It is

well known that scrambling employed in the 2D discretized CSM resembles like random behaviour [30]. The proposed 2D discretized CSM cipher with the OFB and CBC employs IV as a primary key. The IV should be random to be resistant against various types of brute force attacks. The utilized XOR among the IV fragments and data block fragments modifies pixel values, making the proposed 2D discretized CSM cipher with the ECB, OFB and CBC like a standard 3-D map. The proposed 2D discretized CSM cipher with the ECB, OFB and CBC also employs a secondary key, which is utilized in the 2D discretized CSM to shuffle pixels.

Finally, the proposed 2D discretized CSM cipher with the ECB, OFB and CBC image cryptosystem depends on segmenting the images to be cipher into various segments. The segment size in bits considered of a significant factor for affecting the performance of the proposed 2D discretized CSM cipher with the ECB, OFB and CBC. The segment size impact on cipher quality is examined in details in the experimental results part. Since the Section IV provides an equivalent bits size as plaintext segment size, increasing the segment size will result in increasing the security. In addition, the proposed 2D discretized CSM cipher with the ECB, OFB and CBC has the ability for encrypting digital images of any size after splitting them into smaller segments.

V. CIPHER QUALITY KEY INDICATORS

The cipher quality testing is very important for image cipher. Visual encryption quality is not sufficient for this test. So, there is a need for mathematical cipher quality key indicator metrics. Here, four cipher quality indicators will be considered to assess and compare the effectiveness of the proposed 2D discretized CSM cipher with the ECB, OFB and CBC. These quality key indicator metrics include correlation coefficient, irregular deviation, and deviations of histogram. In addition, two other quality key indicator metrics are also considered to assess cipher quality; histogram uniformity, and computational time [31-40].

A. The Correlation Coefficient

The correlation coefficient may be considered as a significant estimation for assessing the ciphering quality of any image cipher. As the correlation coefficient becomes near zero, the performance of the image cipher becomes good [31-32]. The correlation coefficient can be estimated as follows [31-32]:

$$CC = \frac{\text{cov}(x,y)}{\sqrt{D(x)}\sqrt{D(y)}} \quad (9)$$

where x and y denotes the pixels intensity levels at the same location in both the plainimage and ciphered image. The definitions for the covariance, standard deviation and mean are given below as follows:

$$\text{cov}(x,y) = \frac{1}{L} \sum_{i=1}^L (x(i) - E(x))(y(i) - E(y)) \quad (10)$$

$$D(x) = \frac{1}{L} \sum_{i=1}^L (x(i) - E(x))^2 \quad (11)$$

$$D(y) = \frac{1}{L} \sum_{i=1}^L (y(i) - E(y))^2 \quad (12)$$

where, L denotes the pixels number. As the correlation becomes near zero, the better the image cipher quality.

B. Histogram Distribution

The cipherimage histogram distribution can be utilized as an indicator for assessing the quality of the proposed 2D CSM cipher with the ECB, OFB and CBC. As the cipherimage has a uniform histogram distribution, the proposed 2D CSM cipher with the ECB, OFB and CBC has a good ciphering quality.

C. The Irregular Deviation

The irregular may be employed for assessing the ciphering quality in terms of how much it can reduce the deviation to be near the histogram of an ideally ciphered image [33-35]. The process of estimating the irregular deviation starts by calculating the absolute deviation among the plainimage and the enciphered image. After that it calculates the histogram of the resulted absolute deviation. Then, calculate the mean histogram of the resulted absolute deviation. Finally, calculate the histogram deviation absolute mean value.

The irregular deviation can be computed as given below:

$$D_I = \frac{\left| \sum_{i=0}^{255} H(i) - M_H \right|}{M \times N} \quad (13)$$

As the irregular deviation becomes low, the performance of the image cipher becomes good.

D. The Deviation of Histogram

The deviation of histogram can be employed for assessing the ciphering quality in terms of how much it can magnify the difference among the plainimage and the enciphered image [36-38]. The process of estimating the deviation of histogram starts by calculating the histogram of the plainimage and the enciphered image. After that, calculate the absolute deviation among histogram of the plainimage and the enciphered image. Finally, compute the curve area beyond the absolute deviation divided by the total image area as follows [36-38]:

$$D_H = \frac{\left(\frac{d_0 + d_{255}}{2} + \sum_{i=1}^{254} d_i \right)}{M \times N} \quad (14)$$

where d_i denotes is the absolute difference curve magnitude at intensity level i , M and N denote the image dimensions. As the deviation of histogram becomes high, the performance of the image cipher becomes good [25].

E. The Impact of Noise

Noise immunity demonstrates the cipher capability to withstand and against the noise. To examine and measure the impact of noise on the proposed 2D discretized CSM cipher with the ECB, OFB and CBC, noises of various SNRs are added to cipherimages, and after that the deciphering procedure is applied. If the resulted deciphered image is very close to its corresponding plainimage, it could mean that the proposed 2D

discretized CSM cipher with the ECB, OFB and CBC has a capability to resist the noise [39-42]. This proximity can be ensured numerically or visually using the correlation coefficients and the PSNR of the deciphered image, which can be denoted as follows [39-42]:

$$PSNR = 10 \times \log_{10} \left(\frac{M \times N \times 255^2}{\sum_{m=1}^M \sum_{n=1}^N (f(m,n) - f_d(m,n))^2} \right) \quad (15)$$

where $f(m,n)$ denotes the plainimage and $f_d(m,n)$ denotes its corresponding deciphered image.

VI. EXPERIMENTAL TESTS AND DISCUSSION

In experimental tests, test experiments were employed to investigate and examine the proposed 2D discretized CSM cipher with the ECB, OFB and CBC. With respect to the proposed 2D discretized CSM cipher with the ECB, OFB and CBC, the IV is employed as a portion of the enciphered Pirate image, and has an equivalent size with respect to the chosen segment size. Various segments of different sizes were examined in testing the proposed 2D discretized CSM cipher with the ECB, OFB and CBC as shown below:

- 1) $S_1 = IV = 4 \times 4$ with IV as a portion of the enciphered Pirate image.
- 2) $S_2 = IV = 8 \times 8$ with IV as a portion of the enciphered Pirate image.
- 3) $S_3 = IV = 16 \times 16$ with IV as a portion of the enciphered Pirate image.
- 4) $S_4 = IV = 32 \times 32$ with IV as a portion of the enciphered Pirate image.
- 5) $S_5 = IV = 64 \times 64$ with IV as a portion of the enciphered Pirate image.
- 6) $S_6 = IV = 128 \times 128$ with IV as a portion of the enciphered Pirate image.
- 7) $S_7 = IV = 256 \times 256$ with IV as a portion of the enciphered Pirate image.

The enciphered Pirate images using the proposed 2D discretized CSM cipher with the ECB, OFB and CBC and various segment sizes are depicted in Fig. 2. It is clearly noted from Fig. 2 that the proposed 2D discretized CSM cipher with the OFB and CBC has a better performance than the proposed 2D discretized CSM cipher with ECB especially with small segment sizes. Also, with increasing the segment size, the proposed 2D discretized CSM cipher with the OFB and CBC has a good performance in hiding all the details of the enciphered images.

The histograms distribution of Pirate plainimage and its enciphered image using the 2D discretized CSM image cipher with the ECB, OFB and CBC are depicted in Fig. 3. It is clearly noted from Fig. 3 that the 2D discretized CSM image cipher with ECB does not provide histogram uniformity and it has the same histogram of the original Pirate plainimage. This is due to the fact that the 2D discretized CSM image cipher with ECB performs just permutation which does not change the histograms of the encrypted images which may be considered

as a basic weakness. Finally, it is clear from Fig. 3 that the 2D discretized CSM image cipher with the OFB and CBC can provide histogram uniformity.

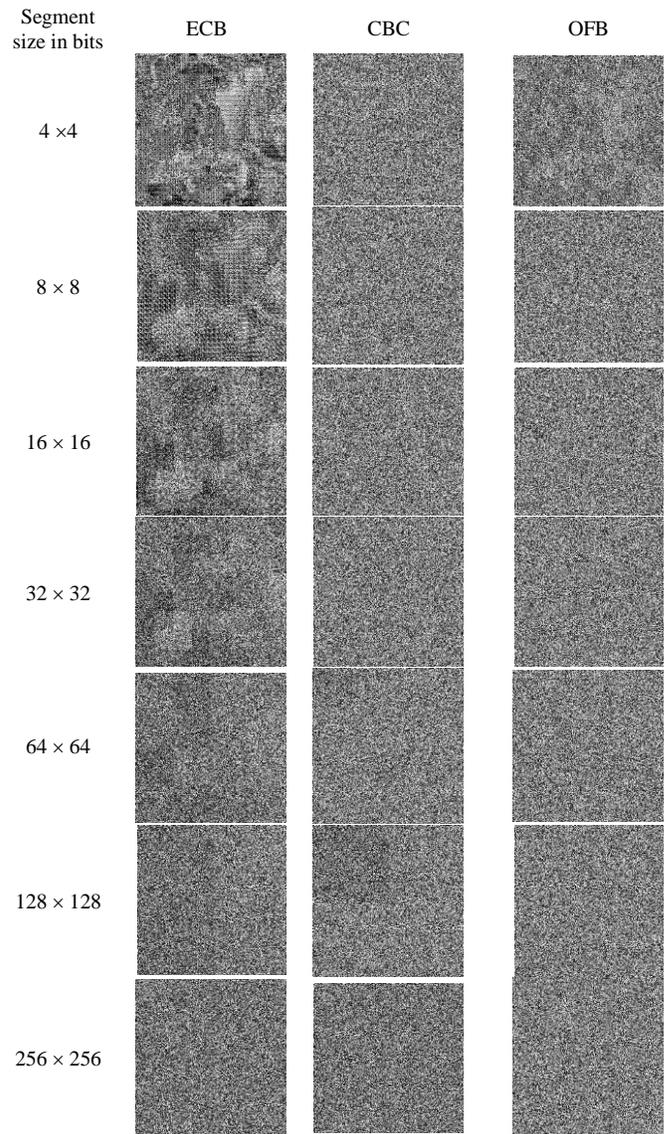


Fig. 2. Enciphered Pirate Images using the Proposed 2D Discretized CSM Cipher with the OFB and CBC with Various Segment Sizes.

Tables I, II, and III illustrates the numerical estimations of the evaluated key performance metrics like correlation coefficient (CC), irregular deviation (ID) and maximum deviation (MD) for the proposed 2D discretized CSM cipher with the ECB, OFB and CBC and various segment sizes.

The CC outcomes results listed in Table I demonstrated that the proposed 2D discretized CSM cipher with the OFB and CBC has lower CC values than in the proposed 2D discretized CSM cipher with the ECB. Also, with increasing the segment size, the CC values of the proposed 2D discretized CSM cipher with the ECB, OFB and CBC becomes near the zero value.

The ID outcomes results listed in Table II demonstrated that the proposed 2D discretized CSM cipher with the OFB and CBC has lower ID values than in the proposed 2D discretized

CSM cipher with the ECB. Also, with increasing the segment size, the ID values of the proposed 2D discretized CSM cipher with the ECB, OFB and CBC decrease.

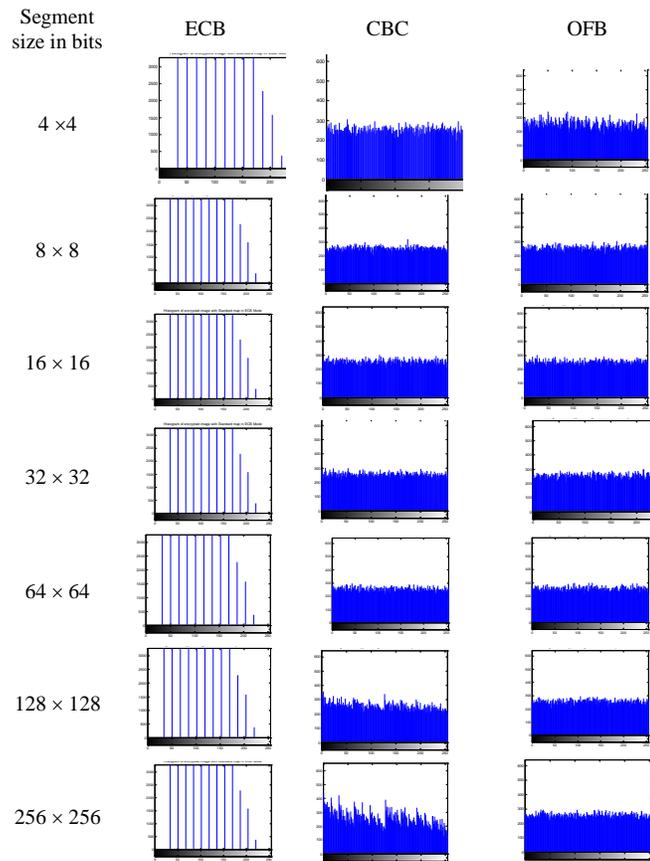


Fig. 3. Histogram of Enciphered Pirate Images using the Proposed 2D Discretized CSM Cipher with the OFB and CBC with Various Segment Sizes.

The MD outcomes results are listed in Table III and the outcomes results demonstrated that the 2D discretized CSM image cipher with ECB provides zero MD values. This is due to the fact that the 2D discretized CSM image cipher with ECB performs just permutation which does not change the histograms of the encrypted images.

Also, the proposed 2D discretized CSM cipher with the OFB and CBC has better MD values than in the proposed 2D discretized CSM cipher with the ECB.

TABLE I. CC OF ENCIPHERED PIRATE IMAGES USING THE PROPOSED 2D DISCRETIZED CSM CIPHER WITH THE ECB, OFB AND CBC WITH VARIOUS SEGMENT SIZES

Segment size in bits	ECB	CBC	OFB
4 × 4	0.1715	-0.0043	0.1156
8 × 8	0.1307	-0.0075	-0.0159
16 × 16	0.1326	-0.00092	-0.0198
32 × 32	0.0765	0.0028	-0.0105
64 × 64	0.0497	-0.0054	0.0044
128 × 128	0.0351	0.0233	0.0028
256 × 256	0.0086	0.0066	-0.0086

TABLE II. ID OF ENCIPHERED PIRATE IMAGES USING THE PROPOSED 2D DISCRETIZED CSM CIPHER WITH THE OFB AND CBC WITH VARIOUS SEGMENT SIZES

Segment size in bits	ECB	CBC	OFB
4 × 4	0.7891	0.6667	0.7489
8 × 8	0.7300	0.6655	0.6677
16 × 16	0.7400	0.6676	0.6647
32 × 32	0.7130	0.6715	0.6655
64 × 64	0.7038	0.6679	0.6690
128 × 128	0.6984	0.6844	0.6711
256 × 256	0.6840	0.6837	0.6682

TABLE III. MD OF ENCIPHERED PIRATE IMAGES USING THE PROPOSED 2D DISCRETIZED CSM CIPHER WITH THE OFB AND CBC WITH VARIOUS SEGMENT SIZES

Segment size in bits	ECB	CBC	OFB
4 × 4	0	1.8963	1.8920
8 × 8	0	1.8978	1.9002
16 × 16	0	1.8988	1.8988
32 × 32	0	1.8980	1.8986
64 × 64	0	1.8934	1.8970
32 × 128	0	1.9020	1.8977
256 × 256	0	1.8992	1.8969

To examine the impact of the noise for the proposed 2D discretized CSM cipher with the ECB, OFB and CBC, the additive white Gaussian noise (AWGN) of SNR equals to 5 dB is summed to the enciphered Pirate image, and the deciphering procedure is applied. The deciphering outcome results of the proposed 2D discretized CSM cipher with the ECB, OFB and CBC and various segment sizes are depicted in Fig. 4. It is clearly noted from Fig. 4 that the proposed 2D discretized CSM cipher with the OFB has a better performance than the proposed 2D discretized CSM cipher with the ECB and CBC with small segment sizes. Also, the proposed 2D discretized CSM cipher with the OFB is more resistant to noise than the proposed 2D discretized CSM cipher with the ECB and CBC.

In addition, the OFB and CBC operation modes provide approximately equivalent performance in the existence of noise. Finally, the segment size has no impact on noise resistance of the proposed 2D discretized CSM cipher with the ECB, OFB and CBC.

Table IV and Table V illustrate the numerical estimations of the evaluated key performance metrics like PSNR and CC for the proposed 2D discretized CSM cipher with the ECB, OFB and CBC and various segment sizes. These PSNR and CC numerical key performance indicator values are estimated in the AWGN existence of SNR equals 5 dB. The PSNR and CC results listed in Table IV and Table V demonstrated that the proposed 2D discretized CSM cipher with the OFB has a better PSNR and CC than the proposed 2D discretized CSM cipher with the ECB and CBC.

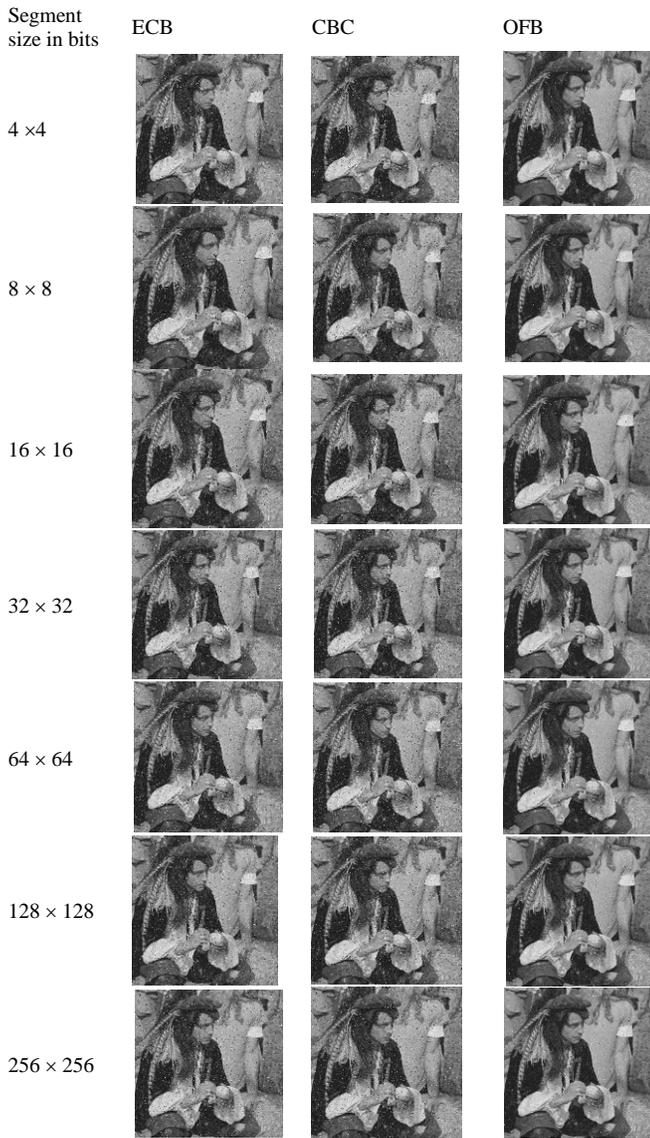


Fig. 4. Deciphered Pirate Images of the Proposed 2D Discretized CSM Cipher with the OFB and CBC with Various Segment Sizes and SNR=5dB.

TABLE IV. NUMERICAL PSNR VALUES OF THE 2D DISCRETIZED CSM IMAGE CIPHER, AND THE PROPOSED 2D DISCRETIZED CSM CIPHER WITH THE OFB AND CBC AND VARIOUS SEGMENT SIZES

Segment size in bits	ECB	CBC	OFB
4 × 4	20.3094	20.5365	29.4117
8 × 8	21.3683	20.6225	29.2111
16 × 16	21.4050	20.6223	29.1333
32 × 32	21.1286	20.5498	29.2305
64 × 64	21.1369	20.4960	29.5665
32 × 128	21.2714	20.8906	28.9257
256 × 256	21.3248	21.3151	29.4045

TABLE V. NUMERICAL CC VALUES OF THE 2D DISCRETIZED CSM IMAGE CIPHER, AND THE PROPOSED 2D DISCRETIZED CSM CIPHER WITH THE OFB AND CBC AND VARIOUS SEGMENT SIZES

Segment size in bits	ECB	CBC	OFB
4 × 4	0.9043	0.8862	0.9842
8 × 8	0.9057	0.8894	0.9834
16 × 16	0.9066	0.8892	0.9831
32 × 32	0.9002	0.8868	0.9836
64 × 64	0.9005	0.8856	0.9848
32 × 128	0.9036	0.8953	0.9824
256 × 256	0.9044	0.9042	0.9842

Also, the proposed 2D discretized CSM cipher with the OFB is more resistant to noise than the proposed 2D discretized CSM cipher with the ECB and CBC. Finally, it can be confirmed and ensured that the proposed 2D discretized CSM cipher with the OFB can provide a better trade-off among the security level and noise immunity.

VII. CONCLUSION

This paper introduces a 2D discretized CSM cipher with the ECB, OFB and CBC that depends on dividing the plainimage to be enciphered into segments and enciphering each segment with the 2D discretized CSM cipher with the ECB, OFB and CBC. The proposed 2D discretized CSM cipher with the OFB provides a better trade-off among the high noise resistivity and high security level. The tests demonstrated that the 2D discretized CSM cipher with the OFB and CBC also achieve a uniform histogram distribution that cannot be achieved using the proposed 2D discretized CSM cipher with the ECB. The 2D discretized CSM is compared in different modes of operation. The outcomes demonstrated that the proposed 2D discretized CSM with the OFB and CBC has a high security. Finally, the proposed 2D discretized CSM cipher with the OFB has good noise immunity than the proposed 2D discretized CSM cipher with the ECB and CBC.

ACKNOWLEDGMENT

This study was funded by the Deanship of Scientific Research, Taif University Researchers Supporting Project number (TURSP-2020/98), Taif University, Taif, Saudi Arabia.

REFERENCES

- [1] O. S. Faragallah, A. Afifi, W. El-Shafai, H. S. El-sayed, M. A. AlZain, J. F. Al-Amri, and F. E. Abd El-Samie, "Efficiently encrypting color images with few details based on RC6 and different operation modes for cybersecurity applications," *IEEE Access*, vol. 8, pp. 103200-103218, 2020.
- [2] O. S. Faragallah, A. I. Sallam and H. S. El-Sayed, "Utilization of HEVC ChaCha20-based selective encryption for secure telehealth video conferencing," *Computers, Materials & Continua*, vol. 70, pp. 831-845, 2022.

- [3] O. S. Faragallah, H. S. El-sayed, A. Afifi, W. El-Shafai, "Efficient and secure opto-cryptosystem for color images using 2D logistic-based fractional Fourier transform," *Optics and Lasers in Engineering*, vol. 137, 106333, 2021.
- [4] C. E. Shannon, "Communication theory of secrecy system," *Bell Syst. Tech. J.*, vol. 28, pp. 656–715, 1949.
- [5] Z. Liu, C. Guo, J. Tan, W. Liu, J. Wu, Q. Wu, L. Pan, S. Liu, "Securing color image by using phase-only encoding in Fresnel domains," *Opt. and Lasers in Eng.*, vol. 68, pp. 87-92, 2015.
- [6] X. W. Li, Q. Wang, S. Kim, and I. Lee, "Encrypting 2D/3D image using improved lensless integral imaging in Fresnel domain," *Opt. Commun.*, vol. 381, pp. 260-270, 2016.
- [7] [19] S. holami, K. Jaferzadeh, and S. Shin, "An efficient image-based verification scheme by fusion of double random phase encoding and dynamic chaotic map," *Multimedia Tools and Applications*, vol. 78, pp. 25001–25018, 2019.
- [8] E. Alvarez, A. Fernández, P. García, J. Jiménez, and A. Marcano, "A new approach to chaotic encryption" *Physics Letters A*, vol. 26, pp. 373-375, 1999.
- [9] O. S. Faragallah, A. Afifi, I. F. Elashry, E. A. Naeem H. M. El-Hoseny, H. S. El-sayed, and A. M. Abbas, "Efficient optical double image cryptosystem using chaotic mapping-based Fresnel transform," *Optical and Quantum Electronics*, vol. 53, pp. 1-26, 2021.
- [10] Z. Hua and Y. Zhou, "Image encryption using 2D logistic-adjusted-Sine map," *Information Sciences*, vol. 339, pp. 237-253, 2016.
- [11] K. Wang, W. Pei, and L. Zou "Security of public key encryption technique based on multiple chaotic system" *Physics Letters A*, vol. 360, pp. 259-262, 2006.
- [12] I. F. Elashry, W. El-Shafai, E. S. Hasan, S. El-Rabaie, A. M. Abbas, F. E. Abd El-Samie, H. S. El-sayed, and O. S. Faragallah, "Efficient chaotic-based image cryptosystem with different modes of operation," *Multimedia Tools and Applications*, vol. 79, pp. 20665-20687, 2020.
- [13] P. Zhen, G. Zhao, L. Min, and X. Jin, "Chaos-based image encryption scheme combining DNA coding and entropy," *Multimed. Tools Appl.*, vol. 75, pp. 6303-6319, 2016.
- [14] J. S. Khan and J. Ahmad, "Chaos based efficient selective image encryption," *Multidim. Syst. Sign. Process.*, vol. 30, no. 2, pp. 943–961, 2019.
- [15] Y. Luo, J. Yu, W. Lai, and L. Liu, "A novel chaotic image encryption algorithm based on improved baker map and logistic map," *Multimed. Tools Appl.*, vol. 78, pp. 22023-22043, 2019.
- [16] H. Chen, C. Tanougast, Z. Liu, W. Blondel, and B. Hao, "Optical hyperspectral image encryption based on improved Chirikov mapping and gyration transform," *Optics and Lasers in Engineering*, vol. 107, pp. 62-70, 2018.
- [17] G. Hu, D. Xiao, Y. Zhang, and T. Xiang, "An efficient chaotic image cipher with dynamic lookup table driven bit-level permutation strategy," *Nonlinear Dynamics*, vol. 87, no. 2, pp. 1359-1375, 2017.
- [18] J. Thomas, "Individual cyber security: Empowering employees to resist spear phishing to prevent identity theft and ransomware attacks," *International Journal of Business Management*, vol. 13, no. 6, pp. 1-24, 2018.
- [19] D. Zhang, "Big data security and privacy protection," *Proc. of 8th IEEE International Conference on Management and Computer Science (ICMCS 2018)*, pp. 275-278, Atlantis Press, October 2018.
- [20] A. Belazi, M. Khan, A. A. Abd El-Latif, and S. Belghith, "Efficient cryptosystem approaches: S-boxes and permutation–substitution-based encryption," *Nonlinear Dynamics*, vol. 87, no. 1, pp. 337-361, 2017.
- [21] S. Lian, G. Sun, and Z. Wang, "A block cipher based on a suitable use of chaotic Standard map," *Chaos, Solutions and Fractals*, vol. 26, pp. 117-129, 2005.
- [22] G. Hu, D. Xiao, Y. Zhang and T. Xiang, "An efficient chaotic image cipher with dynamic lookup table driven bit-level permutation strategy," *Nonlinear Dynamics*, vol. 87, no. 2, pp. 1359-1375, 2017.
- [23] K. Wong, B. Kwok, and W. Law W, "A Fast Image Encryption Scheme based on Chaotic Standard Map," *Phys. Lett A*, vol. 37, pp. 112-117, 2007.
- [24] O. S. Faragallah, A. I. Sallam and H. S. El-Sayed, "Visual protection using RC5 selective encryption in telemedicine," *Intelligent Automation & Soft Computing*, vol. 31, pp. 1717-190, 2022.
- [25] J. Kohl, "The use of encryption in Kerberos for network authentication," *Proceedings, Crypto-89*, Springer-Verlag, 1989.
- [26] M. Dworkin, Recommendation for block cipher modes of operation methods and techniques, NIST Special Publication 800-38A, 2001.
- [27] B. Stoyanov and G. Nedzhibov, "Symmetric key encryption based on rotation-translation equation," *Symmetry*, vol. 12, pp. 1-12, 2020.
- [28] A. Arab, M. J. Rostami and B. Ghavami, "An image encryption method based on chaos system and AES algorithm," *The Journal of Supercomputing*, vol. 75, pp. 6663–6682, 2019.
- [29] X. J. Tong and M. G. Cui, "Image encryption with compound chaotic sequence cipher shifting dynamically, IVC," vol. 26, pp. 843-850, 2008.
- [30] M. H. Abood, "An efficient 3DV frame cryptography using hash-LSB steganography with RC4 and pixel shuffling encryption algorithms," *Annual Conference on New Trends in Information & Communications Technology Applications (NTICT)*, pp. 86-90, 2017.
- [31] A. Abukari, E. Bankas, and M. Iddrisu, "A secured video conferencing system architecture using a hybrid of two homomorphic encryption schemes: a case of zoom," *International Journal of Engineering Research & Technology (IJERT)*, vol. 9, pp. 237-240, 2020.
- [32] O. S. Faragallah, H. S. El-sayed, A. Afifi, and S. F. El-Zoghdy, "Small details gray scale image encryption using RC6 block cipher," *wireless Personal Communications*, vol. 118, no. 2, pp. 1559-1589, 2021.
- [33] O. S. Faragallah, M. A. AlZain, H. S. El-sayed, J. F. Al-Amri, W. El-Shafai, A. Afifi, E. A. Naeem, and B. Soh, "Secure color image cryptosystem based on chaotic logistic in the FrFT domain," *Multimedia Tools and Applications*, vol. 79, pp. 2495-2519, 2020.
- [34] O. S. Faragallah, A. Afifi, W. El-Shafai, H. S. El-sayed, E. A. Naeem, M. A. AlZain, J. F. Al-Amri, B. Soh, and F. E. Abd El-Samie, "Investigation of chaotic image encryption in spatial and FrFT domains for cybersecurity applications," *IEEE Access*, vol. 8, pp. 42491-42503, 2020.
- [35] Z. Xiong, K. Ramchandran, M. T. Orchard, and Ya-Qin Zhang, "A Comparative study of DCT- and wavelet-based image coding," *IEEE transactions on circuits and systems for video technology*, vol. 9(5), pp. 352-367, August 1999.
- [36] O. S. Faragallah, W. El-Shafai, A. Afifi, I. Elashry, M. A. AlZain, J. F. Al-Amri, B. Soh, H. M. El-Hoseny, H. S. El-Sayed, and F. E. Abd El-Samie, "Efficient three-dimensional video cybersecurity framework based on double random phase encoding," *Intelligent Automation & Soft Computing*, vol. 28, pp. 353-367, 2021.
- [37] A. Sallam, E. EL-Rabaie, and O. S. Faragallah, "HEVC selective encryption using RC6 block cipher technique," *IEEE Transactions on Multimedia*, vol. 20, no. 7, pp. 1636-1644, 2018.
- [38] A. M. Hemdan, O. S. Faragallah, O. Elshakankiry, and A. Elmalaway, "A fast hybrid image cryptosystem based on random generator and modified logistic map," *Multimedia Tools and Applications*, vol. 78, no. 12, pp. 16177-16193, 2019.
- [39] S. Sun, "A Novel Hyperchaotic Image Encryption Scheme Based on DNA Encoding, Pixel-Level Scrambling and Bit-Level Scrambling," *IEEE Photonics Journal*, vol. 10, pp. 1-14, 2018.
- [40] O. S. Faragallah, W. El-Shafai, A. I. Sallam, I. Elashry, E. M. EL-Rabaie, A. Afifi, M. A. AlZain, J. F. Al-Amri, F. E. Abd El-Samie, and H. S. El-sayed, "Cybersecurity framework of hybrid watermarking and selective encryption for secure HEVC communication," *Journal of Ambient Intelligence and Humanized Computing*, vol. 13, pp. 1215-1239, 2022.
- [41] G. Hu, D. Xiao, Y. Zhang, and T. Xiang, "An efficient chaotic image cipher with dynamic lookup table driven bit-level permutation strategy," *Nonlinear Dynamics*, vol. 87, pp. 1359-1375, 2017.
- [42] O. S. Faragallah, A. Afifi, H. S. El-sayed, M. A. AlZain, J. F. Al-Amri, F. E. Abd El-Samie, and W. El-Shafai, "Efficient HEVC integrity verification scheme for multimedia cybersecurity applications," *IEEE Access*, vol. 8, pp. 154112-154135, 2020.

Flower Pollination Algorithm for Feature Selection in Tweets Sentiment Analysis

Muhammad Iqbal Abu Latiffi¹, Mohd Ridzwan Yaakub², Ibrahim Said Ahmad³

Center for Artificial Intelligence Technology, Faculty of Information Science and Technology

Universiti Kebangsaan Malaysia, Bangi Malaysia^{1,2}

Faculty of Computer Science and Information Technology, Bayero University Kano, Kano Nigeria³

Abstract—Text-based social media platforms have developed into important components for communication between customers and businesses. Users can easily state their thoughts and evaluations about products or services on social media. Machine learning algorithms have been hailed as one of the most efficient approaches for sentiment analysis in recent years. However, as the number of online reviews increases, the dimensionality of text data increases significantly. Due to the dimensionality issue, the performance of machine learning methods has been degraded. However, traditional feature selection methods select attributes based on their popularity, which typically does not improve classification performance. This work presents a population-based metaheuristic for feature selection algorithms named Flower Pollination Algorithms (FPA) because of their propensity to accept less optimum solutions and avoid getting caught in local optimum solutions. The study analyses tweets from Kaggle first with the usual Term Frequency-Inverse Document Frequency statistical weighting filter and then with the FPA. Four baseline classifiers are used to train the features: Naive Bayes (NB), Decision Tree (DT), Support Vector Machine (SVM), and k-Nearest Neighbor (kNN). The results demonstrate that the FPA outperforms alternative feature subset selection algorithms. For the FPA, an average improvement in accuracy of 2.7% is seen. The SVM achieves a better accuracy of 98.99%.

Keywords—Sentiment analysis; metaheuristic algorithm; flower pollination algorithm; machine learning; feature selection

I. INTRODUCTION

The introduction of the second generation of the web has resulted in the development of more interactive websites where users will play a significant role by contributing their thoughts, suggestions, and comments via the website's services. Large-scale communication, which is not limited to comments on news, services, or even goods or products, can also be used to facilitate conversation amongst website users within social networks. Social networks allow users to contact one another, engage, and collaborate to become content creators on a single platform known as social media. Through social media, they can provide views, comments, and experiences on an issue [1]–[6].

There are many web users, and there is also a large amount of information sharing in the form of reviews, which makes analyzing the reviews manually tricky and time-consuming. Sentiment Analysis (SA) has been introduced as a practical approach to determining the sentiments included in web materials to tackle this issue [7]–[10].

SA is a type of text classification problem that involves subjective statements. Another name for sentiment analysis is Opinion Mining (OP), where an opinion or comment is processed to determine the consumer's perception of a matter or issue. SA is one of the areas of data mining and text mining that falls under the category of web content mining techniques [8], [11]–[13]. It is defined as the computer study of public perceptions, beliefs, and moods around a specific topic [14]. It is a subset of natural language processing jobs that assesses a large amount of viewer content on social media, blog posts, e-commerce portals, and other user-editable online forums.

One of the primary challenges associated with sentiment analysis is the preprocessing phase of the text, in which they must handle user feedback on social media, webpages, and blog posts that contain irrelevant and noisy content. [15] studied the effects of text preprocessing in sentiment analysis. They found inconsistent text classification results, which is likely due to inefficient text preprocessing methods. Most researchers use only a fraction of the text preprocessing techniques [6], [16], [17]. In addition to improvements in the feature selection phase, the text preprocessing phase also plays a significant role in improving text or sentiment classification performance [2], [18]. According to [18], the habit will increase the results of text classification accuracy.

The magnitude of the feature dimension for text parsing is an additional key difficulty in sentiment analysis. Bag-of-Words is commonly used to represent document text in machine learning algorithms for sentiment categorization [19], [20]. Words in a text document contain feature vectors with large dimensions. As a result, the feature selection process focuses on selecting the optimal subset of features from a large-dimensional feature size. This is accomplished by removing cluttered and irrelevant features without altering the original data [14], [21].

According to [19], it is vital to select an optimal subset of features that represents the actual feature subsets to reduce feature size and increase classification accuracy. An optimization algorithm was utilized as the strategy for selecting features [14]. Previous research [22]–[25] has demonstrated that an optimization algorithm strategy for feature selection can handle feature selection and feature reduction issues in large amounts of data containing noise, redundant, and inaccurate information.

Flower Pollination Algorithm (FPA) is one of the algorithms inspired by nature. This algorithm has advantages in

terms of performance when it is applied in improving performance in various optimization issue problems as well as having few parameters [22], [26], [27]. According to [22], the FPA is a flexible and easily adaptable optimization method. However, research on the use of FPA in feature selection problems in sentiment analysis has not yet been conducted. So, this is necessary to study and further improve the performance of FPA for feature selection problems. Therefore, the development of feature selection methods that use FPA is expected to be able to select a more discriminatory feature set and improve the sentiment classification results.

On the other hand, the sentiment classification process is another issue in sentiment analysis technology due to the involvement of the textual data [5]. Furthermore, simple text classification and sentiment classification in text mining are two different things. Text classification in text mining identifies topics found in a data set. The classification of sentiment, on the other hand, is classified in terms of the type of sentiment in the text. When there are strategies to minimize the dimensions in large-sized data sets, sentiment classification performance can be enhanced [10], [28], [29]. This method generally serves to detect and eliminate irrelevant and overlapping data so that the sentiment classification results are meaningful.

The remaining sections of this work are structured as follows: The second section examines previous works on feature selection and the necessary component of FPA. Next, Section 3 provides an overview of the technique applied in this study, while Section 4 details the FPA algorithm proposed for feature selection in sentiment analysis. The findings of the experiments are discussed in Section 5, and the study is concluded in Section 6.

II. RELATED WORK

Different approaches for selecting features have been proposed and developed, and they are explored by researchers. Metaheuristic techniques improve solution performance over and over again [30]. These metaheuristic techniques are based on observations of natural phenomena such as ant colony optimization, bat algorithms, and flower pollination optimization. This metaheuristic is then used in the bandage feature selection method. This method has been gaining more attention recently [31] as it attempts to produce better solutions by applying knowledge gained from natural solutions. The wrapper approach in feature selection evaluates the quality of the selected features based on their classification performance. The wrapping method has two main steps: (1) finding a subset of features and (2) evaluating the selected features. Both of these steps will be repeated until the stop criteria are met. It starts with the production of a subset, and then the classification will evaluate the subset produced.

Previous studies proposed feature selection methods based on particle clustering optimization and the nearest k-neighbor classification [24]. In addition, based on an analysis from [32] also proposed a hybrid approach of feature selection based on genetic algorithms for the classification of Arabic texts using a wrapper model. In the first step, six feature evaluation methods are used simultaneously to select a subset of features. Then an

enhanced genetic algorithm is used to optimize the selected subset.

The technique of selecting the optimal subset of features using firefly's algorithm for sentiment analysis problems has also proven the ability of this metaheuristic algorithm [16]. While in a study conducted by [17] where researchers improved the whale optimization algorithm for feature selection of two types of sentiment analysis data, namely in Arabic and English. Moreover, the hybridization between the ant colony optimization algorithm and k-Nearest Neighbors has successfully improved the sentiment classification results as it has been applied to the feature selection process [6]. These studies report that their method is more effective compared to the use of filter methods.

The FPA is an algorithm inspired by nature that imitates the basic pollination activity of flowers. In [33], four rules are idealized. Rule 1 of global pollination incorporates biotics and cross-pollination, with the pollinating agent transporting pollen according to the Lévy flight. Rule 2 requires abiotic and self-pollination for local pollination. Next, for Rule 3, the flower constant can be interpreted as the reproduction probability proportional to the degree of resemblance between two flowers. Rule 4 is the exchange probability p [0, 1], which can be regulated by external factors such as wind between local and global pollination. Local pollination accounted for a sizeable proportion of the total pollination activity.

FPA was successfully adapted for several domains of optimization problems [22]. For the field of electrical and power generation, [34] has introduced a Modified FPA (MFPA) that employs dynamic switching probabilities, the application of Real Coded Genetic Algorithm (RCGA) mutations for global and local searches, and differentiation between searches temporary localization and optimal solutions. The MFPA was subsequently assessed for ten power system benchmarks, and the experimental results revealed lesser fuel costs than the FPA. Another study by [35] presented MFPA to assess the fuel funding and time required to get to the globally optimal solution. The Institute of Electrical and Electronics Engineers 30 (IEEE 30) bus test system demonstrated that MFPA outperformed FPA and metaheuristic algorithms.

Next, research involving FPA was conducted on the signal and image processing domains. The Binary Flower Pollination Algorithm (BFPA) has been implemented to solve the challenge of lowering the number of sensors necessary to identify individuals' electroencephalogram signals [36]. BFPA is used to choose the ideal selection of channels that provide maximum accuracy. Based on the Optimum-Path Forest classifier, the findings of the BFPA experiment indicate an identification rate of up to 87 percent. In addition, [37] has done a thorough evaluation of BFPA's performance in solving the Antenna Positioning Problem (APP) area. BFPA was evaluated with real, artificial, and random data of various dimensions and compared with Population-Based Increment Learning (PBIL) and Differential Evolution (DE) algorithms, two efficient APP domain methods. In the sphere of APP, FPA obtains more competitive technical advances than PBIL and DE.

FPA is also not left behind to be applied to the clustering and classification domain. The performance of the modified FPA is tested through a clustering problem. This algorithm was evaluated with several different optimization algorithms, including bat, firefly, and conventional FPA on 10 cluster data sets. From the 10 data sets, 8 were generated from pattern recognition, and 2 were generated artificially. The clustering result is calculated in terms of the value of the objective function and the time taken by the CPU on each run. The distribution length graph illustrates the convergence behavior of the algorithm. The results show that the modified FPA exceeds the comparable algorithm to achieve the best fitness value and reduce CPU processing time [38].

Aside from that, the BFPA was applied to feature selection problems and tested on six data sets, where it outperformed Particle Swarm Optimization (PSO), Harmony Search (HS), and Firefly Algorithm (FA) [39]. BCFA is a hybrid algorithm that combines the Clonal Selection Algorithm (CSA) with FPA to tackle feature selection issues. It was introduced in [40]. Using the Optimum-Path Forest classifier as an objective function with the proposed hybrid algorithm (BCFA) has led to superior performance compared to existing metaheuristics. A new methodology for multi-objective feature selection is based on FPA and rough set theory to identify the optimal classification feature set [41]. This model selects features using the filter and wrapper method. The filter approach is a data-driven methodology, whereas the wrapper method is a classification-based technique. Comparing this method against FPA, PSO, and genetic algorithms, the performance of the suggested method was validated using eight UCI data sets, which revealed that this method is extremely competitive. The addition of FPA to the Ada-Boost algorithm enhances the classification accuracy of text documents during the initial phase of feature selection. In contrast, it is utilized to categorise text materials. Three standard data sets were used to evaluate the performance of the proposed algorithm: CADE 12, WEBKB, and Reuters-21578. The experimental findings demonstrated that the suggested algorithm outperformed Ada-Boost and other algorithms [42].

This paper proposes a metaheuristic approach called the flower pollination algorithm for text feature selection based on the Twitter dataset to increase the accuracy of sentiment classification.

III. METHODOLOGY

In general, the SA methodology is intended to yield the optimal subset of features and the most accurate sentiment classification. As illustrated in Fig. 1, the approach for this study consists of four phases: text preprocessing, feature selection, sentiment classification, testing, assessment, and analysis. The tasks associated with these phases are described in the sections that follow.

A. Phase 1: Text Preprocessing

As tweets are composed by regular people who are not language specialists, the Twitter dataset underwent a cleansing procedure. As a result, misspellings, grammatical errors such as faulty punctuation and capitalization, slang words that do not exist in dictionaries, and abbreviations or acronyms for

common terminology are likely to be present in the datasets. The dataset is therefore subjected to two forms of text preprocessing: linguistic processing and natural language processing. This study's linguistic processing comprises five text processing approaches, including lowercase conversion, removal of '#', '@', and other symbols, and removal of punctuation. The document then undergoes spelling correction, an NLP approach. The preprocessing techniques used in this work are shown in Table I.

B. Phase 2: FPA for Feature Selection

Wind or pollinating agents, such as insects, butterflies, bees, birds, and bats, carry pollen from one flower to another during pollination. Flowering plants have evolved to generate nectar or honey in order to entice pollinators and ensure pollination [43]. In addition, a number of pollination agents and plant species, such as squirrels and ornithophilous flowering plants that are pollinated by birds, constitute a number of floral evolution constants [44]. Based on the main characteristics of pollination, a flower pollination algorithm has been developed by [33] developed an algorithm for flower pollination.

There are two fundamental types of flower pollination: biotic and abiotic processes. Biotic pollination, also known as cross-pollination, is the most common type of pollination and is carried out by pollinating agents like insects, birds, and others. This method of pollination is utilized by over 90 percent of blooming plants. When the pollinating agent travels and even flies at varying speeds, the movement of the pollen is quite remote; such pollination can also be considered global pollination when Lévy Flight rules [33], [45], [46] are used. If pollen is encoded as a solution vector, this operation corresponds to a global search. Abiotic pollination, also known as self-pollination, does not require an external pollinator. Approximately 10 percent of flowering plants utilize this kind of pollination, according to estimates. Since self-pollination and localised pollination are more likely to occur in this way, wind dispersal is an option [43], [44]. Typically, the distance traversed by a local movement is shorter, therefore the search might be deemed local. The aforesaid characteristics were utilised to plan the creation of the flower pollination algorithm (FPA) [33], an optimization algorithm.

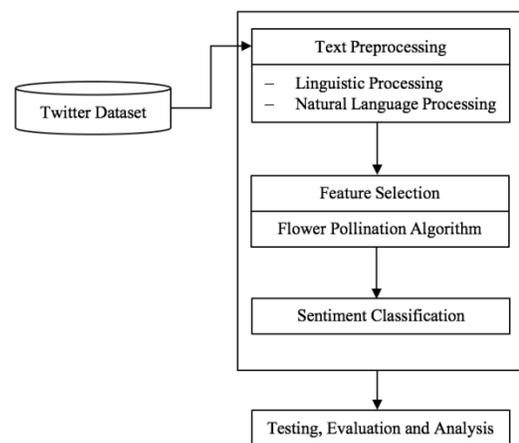


Fig. 1. Methodology for the Study.

TABLE I. TEXT PREPROCESSING AND EXAMPLES

Techniques	Raw	Processed
Conversion to lowercase	@Piwi_47 I hated the da Vinci code, the movie with a passion, it was boring and made me sad!! that a movie blaspheming the name of Jesus is being played world wide \$. #davincicode	@piwi_47 i hated the da vinci code, the movie with a passion, it was boring and made me sad!! that a movie blaspheming the name of jesus is being played world wide \$. #davincicode
Removal of @	@piwi_47 i hated the da vinci code, the movie with a passion, it was boring and made me sad!! that a movie blaspheming the name of jesus is being played world wide \$. #davincicode	i hated the da vinci code, the movie with a passion, it was boring and made me sad!! that a movie blaspheming the name of jesus is being played world wide \$. #davincicode
Removal of punctuation	i hated the da vinci code, the movie with a passion, it was boring and made me sad!! that a movie blaspheming the name of jesus is being played world wide \$. #davincicode	i hated the da vinci code the movie with a passion it was boring and made me sad that a movie blaspheming the name of jesus is being played world wide \$ #davincicode
Removal of #	i hated the da vinci code the movie with a passion it was boring and made me sad that a movie blaspheming the name of jesus is being played world wide \$ #davincicode	i hated the da vinci code the movie with a passion it was boring and made me sad that a movie blaspheming the name of jesus is being played world wide \$
Removal of symbol	i hated the da vinci code the movie with a passion it was boring and made me sad that a movie blaspheming the name of jesus is being played world wide \$	i hated the da vinci code, the movie with a passion it was boring and made me sad that a movie blaspheming the name of jesus is being played world wide
Spelling correction	it was really ironic that he spent the first part of class talking about his own professot at Harvard who was a pompous arrogant ass	it was ironic that he spent the first part of class talking about his professor at Harvard who was a pompous arrogant ass

The development of the FPA is based on the pseudo-code shown in Fig. 2. In general, the development of the FPA consists of three main parts - parameter declaration, initiation, and searching.

```

1  Parameter Declaration
2  Initialize population size randomly
3  Initialize stopping criteria (iteration number)
4  Set the switch probability
5  Set the best solution
6
7  Initiation phase
8  Check if stopping criteria meet
9  {
10
11  Searching Phase
12  Global pollination
13  Choose the best solution
14  Random solution < switch probability
15  Global pollination occurs
16  New solution produced
17  If new solution better
18      Subs the current solution with the new
19  Else
20      New solution rejected
21  Local pollination
22  Choose two random solution
23  Local pollination occurs
24  New solution produced
25  If new solution better
26      Subs the current solution with the new
27  Else
28      New solution rejected
29  Repeat step # 6
30 }
31 Select the optimal solution

```

Fig. 2. Pseudocode for FPA.

1) *Parameter declaration phase:* There are two types of parameters that must be set during the general parameter declaration phase: population size and the number of iterations. The population parameter choice for this experiment is based on a study by [47], which determined that the optimal population size for optimal results is 25. The number of generations is determined based on a study conducted by [33], i.e., the most appropriate number of generations is 100. This number of generations will be the termination criterion for the algorithm.

2) *Initiation phase:* In this initial phase, a subset of solutions will be randomly generated and stored in the form of a one-dimensional array, as in the example shown in Fig. 3. The illustration depicts a solution subset with ten feature attributes labeled F1 through F10. A cell with a value of 1

indicates the attribute of the selected feature, while a cell with a value of 0 indicates the attribute of the unselected feature. Next, all subsets of these generated initial solutions will be evaluated based on the performance of the sentiment classification accuracy assessment and sorted based on the classification accuracy score values obtained. A subset of this initial solution will be used to generate the following generation subset.

3) *Searching phase:* For the searching phase, the algorithm is divided into two parts, namely, global pollination and local pollination. Before proceeding to the search phase, the termination criteria will be reviewed first. If the termination criteria are not met, the search phase will continue. Conversely, suppose the termination criteria have been met. In that case, the subset of features with the best sentiment classification score will be returned and used as a feature list for sentiment classification.

Global pollination begins when a solution is randomly generated. If its value is smaller than the exchange probability, the Levy flight rule will generate a new solution. Existing solutions will be selected along with these new solutions to get better scores. The crossover method in the global pollination method is illustrated in Fig. 4. The global pollination session will result in two new solutions, and both will go through a sentiment classification performance appraisal process. Only new solutions that are capable of producing higher scores than the existing solution score list is accepted into the population, and their position within the population is decided by the scores produced. On the other hand, this new solution will not be accepted if it gets a score that is less than the score of the existing population.

F1	F2	F3	F4	F5	F6	F7	F8	F9	F10
1	0	1	0	1	1	1	1	0	0

Fig. 3. Solution Subset Array.

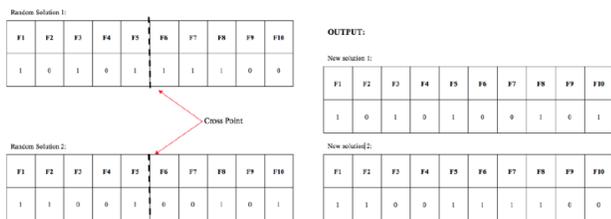


Fig. 4. Global Pollination Process.

On the other hand, for local pollination sessions, it occurs when a solution is randomly generated and if its value is greater than the exchange probability. Therefore, two randomly generated solutions were selected to undergo local pollination. As depicted in Fig. 5, the process of learning between the two candidates in the FPA algorithm occurs via the crossover approach. The outcomes of the local pollination will generate two new solutions, which are then evaluated based on their performance in sentiment classification. If based on the obtained sentiment evaluation score, this new solution is deemed superior to the previous solution, it will be accepted and its population position will be changed. Alternatively, this solution will not be approved if it has a low score among the current solutions.

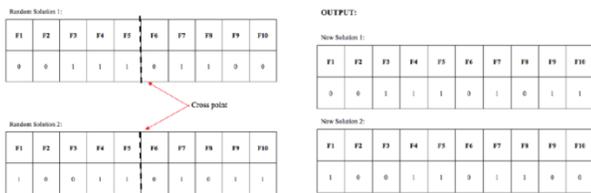


Fig. 5. Local Pollination Process.

These global and local pollination processes will continue until the termination criteria are reached. If the criteria for stopping are not reached, this procedure will be repeated. As illustrated in Fig. 6, the result of this algorithm is a subset of quality and modest characteristics that will be employed in the sentiment classification step.

F1	F2	F3	F4	F5	F6	F7	F8	F9	F10
1	0	1	0	1	1	1	0	0	0

Fig. 6. New Feature Subset that Selected.

C. Phase 3: Sentiment Classification

Classification is one of the key processes in machine learning that refers to determining input data assigned to one of predetermined categories or classes [48]. Classifiers will learn with a learning algorithm or classification algorithm, also known as a classification initiator, a supervised machine learning algorithm. Machine learning algorithms use a set of examples to learn to classify the class label of something that has not been seen or has not been learned. The classifier that has been learned will take the feature value or attribute of an object as input and the class label that has been defined as the output. A set of class labels is defined as part of a problem by the user.

Therefore, the third phase involves four machine learning classifier algorithms to carry out the sentiment classification process by matching a subset of the features generated with the feature information found in the text. The algorithms used are SVM, NB, DT, and kNN.

D. Phase 4: Testing, Evaluation and Analysis

1) *Dataset*: For this study, a benchmark Twitter data set in which has been used by the previous researcher [49] can be obtained from the Kaggle repository. This data set consists of 7086 positive and negative Twitter comments written in English.

2) *Baseline model*: The performance of the FPA feature selection technique was evaluated through a comparison with two baseline algorithms, namely, TF-IDF and Binary Cuckoo Search, by [49]. The evaluation is based on the sentiment classification accuracy from the tweets by using a subset of features selected from phase 3. A subset of these features is obtained upon the text preprocessing process, reduction of feature dimension size, and subsequently feature selection process. This study will not cover the time required for the text preprocessing phase, feature selection, and even classification. Thus, measurements of the time rate will not be performed.

3) *Evaluation*: The accuracy measure was used to evaluate the efficacy of the FPA feature selection technique in getting the optimal subset of features. The evaluation is based on the accuracy of the sentiment classification process outcomes. The classification algorithm generates a confusion matrix, which is used to guide the evaluation process. The confusion matrix displays information about the actual number of classes as well as the number of predictions made by the classification algorithm. True positive (TP) is a condition in which a positive case is successfully classified as positive. True negative (TN) are conditions in which a negative case is successfully classified as negative. A false positive (FP) is a negative case but is misclassified as a positive. False-negative (FN) are positive cases but misclassified as negative ones, as shown in Table II.

The accuracy of the proposed feature selection algorithm was used to evaluate its effectiveness. Accuracy is a simple performance metric derived as the ratio of successfully predicted values to total values. The equation is shown in Equation 1.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

TABLE II. THE CONFUSION MATRIX

		Actual Class	
		Yes	No
Classification Result	Yes	TP	FN
	No	FP	TN

IV. FLOWER POLLINATION FOR FEATURE SELECTION

For this experiment, the evaluation performance was evaluated based on accuracy. These values are compared with the results obtained from baseline algorithms, as presented in Table III.

Table III shows that the FPA feature selection algorithm achieves the highest accuracy value compared to the TF-IDF and Binary Cuckoo Search (BCS). Fig. 7 displays the overall accuracy values for the APB algorithm and TF-IDF and BCS based on the classification algorithm used. Based on Table III, it is found that the accuracy values for TF-IDF for NB, SVM, DT, and k-NN are 87.71%, 89.21%, 89.90%, and 88.42%, respectively. While for Binary Cuckoo Search, the resulting accuracy values are 96.26%, 96.54%, 96.26%, and 95.56% for the NB, SVM, DT, and k-NN, respectively. Based on Fig. 7, the FPA feature selection algorithm has produced the highest accuracy compared to the other two feature selection algorithms. This is where the accuracy values obtained for the NB, SVM, DT, and k-NN are 98.79%, 98.99%, 98.76%, and 98.91%, respectively.

This experiment went through 100 iterations. As shown in Fig. 8, at the 45th iteration, the algorithm is approaching the optimum value at a rapid pace of convergence. This indicates that the algorithm can discover a solution quickly. The increment on accuracy rate then slows until the 75th iteration. This is because this algorithm has consolidated nearly every high feature set. Achieving a subset of features capable of obtaining higher accuracy values got progressively challenging because the current feature set already had a relatively high quality of accuracy values—the accuracy rate value peaks as it reaches the 75th iteration. As a result, until the 100th iteration, this accuracy value remains constant.

TABLE III. ACCURACY PERFORMANCE FOR FPA, TF-IDF AND BCS USING NAÏVE BAYES, SUPPORT VECTOR MACHINE, DECISION TREE AND K-NEAREST NEIGHBOR

Classifiers	FPA	BCS	TF-IDF
Naïve Bayes	98.79	96.26	87.71
Support Vector Machine	98.99	96.54	89.12
Decision Tree	98.76	96.26	89.90

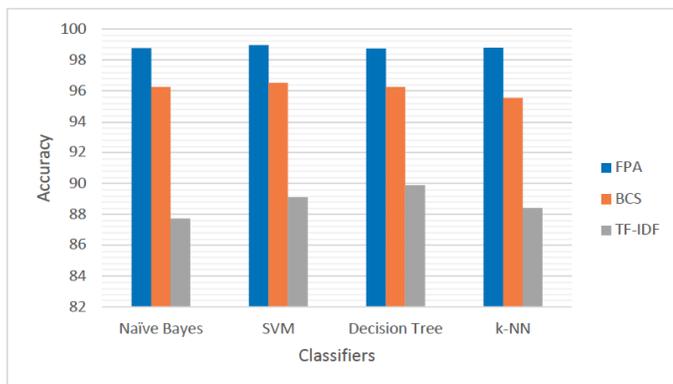


Fig. 7. Comparison of Accuracy for FPA, TF-IDF, and BCS.

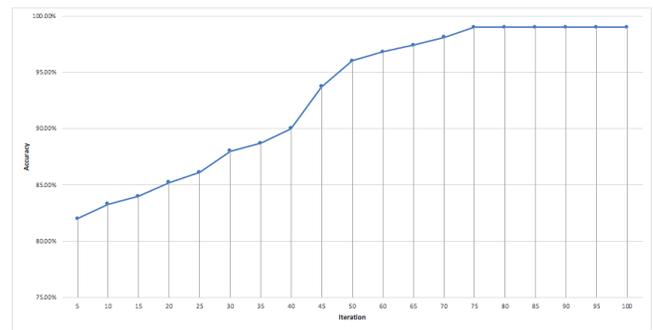


Fig. 8. Classification.

The proposed algorithm for feature selection was able to attain greater accuracy than the two baseline techniques. This result demonstrates that FPA is more effective at extracting features than the baseline techniques.

V. CONCLUSION

This study compares and contrasts the NLP method, spelling correction in text preprocessing techniques, with other conventional text preprocessing techniques. In addition, the use of FPA algorithms for feature selection strategies to enhance the performance of sentiment classification has been proposed. Based on the accuracy results, our approach achieved promising results, confirming that implementing NLP (spelling correction) approach on text preprocessing technique and FPA algorithm on feature selection technique improved sentiment classification performance by 2.68 % compared to the baseline model.

In the future, we would like to employ the proposed technique for solving sentiment analysis tasks on a larger data set containing product review evaluations.

ACKNOWLEDGMENT

The authors gratefully acknowledge Universiti Kebangsaan Malaysia and the Industrial Grant Scheme with Perunding Tamadun Teras Pte Ltd for supporting this research project through grant no. GP-2020-K011466 and TT-2020-008.

REFERENCES

- [1] Dritsas, G. Vonitsanos, and I. E. L. B, “Pre-processing Framework for Twitter,” IFIP Int. Fed. Inf. Process. 2019, vol. 2, pp. 138–149, 2019, doi: 10.1007/978-3-030-19909-8.
- [2] S. R. Priya and M. Devapriya, “The role of pre-processing on unstructured and informal text in diabetic drug related twitter data,” Int. J. Sci. Technol. Res., vol. 8, no. 10, pp. 607–611, 2019.
- [3] L. Zhou et al., “Text preprocessing for improving hypoglycemia detection from clinical notes – A case study of patients with diabetes,” Int. J. Med. Inform., vol. 129, no. January, pp. 374–380, 2019, doi: 10.1016/j.jmedinf.2019.06.020.
- [4] I. S. Ahmad, A. Abu Bakar, M. R. Yaakub, and M. Darwich, “Sequel movie revenue prediction model based on sentiment analysis,” Data Technol. Appl., vol. 54, no. 5, pp. 665–683, 2020, doi: 10.1108/DTA-10-2019-0180.
- [5] M. R. Yaakub, M. I. A. Latiffi, and L. S. Zaabar, “A Review on Sentiment Analysis Techniques and Applications,” IOP Conf. Ser. Mater. Sci. Eng., vol. 551, no. 1, 2019, doi: 10.1088/1757-899X/551/1/012070.
- [6] 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, pp. 133–158, 2019, doi: 10.3233/IDA-173740.

- [7] B. Liu, "Sentiment Analysis and Opinion Mining," Morgan Claypool Publ., no. May, pp. 1–108, 2012, doi: 10.2200/S00416ED1V01Y201204HLT016.
- [8] M. T. Khan, M. Durrani, A. Ali, I. Inayat, S. Khalid, and K. H. Khan, "Sentiment analysis and the complex natural language," *Complex Adapt. Syst. Model.*, vol. 4, no. 1, 2016, doi: 10.1186/s40294-016-0016-9.
- [9] M. R. Yaakub, Y. Li, and J. Zhang, "Integration of Sentiment Analysis into Customer Relational Model: The Importance of Feature Ontology and Synonym," *Procedia Technol.*, vol. 11, pp. 495–501, Jan. 2013, doi: 10.1016/J.PROTCY.2013.12.220.
- [10] I. S. Ahmad, A. A. Bakar, M. R. Yaakub, and M. Darwich, "Beyond Sentiment Classification : A Novel Approach for Utilizing Social Media Data for Business Intelligence," *Int. J. Adv. Comput. Sci. Appl.*, vol. 11, no. 3, pp. 437–441, 2020.
- [11] M. K. Sohrabi and F. Hemmatian, "An efficient preprocessing method for supervised sentiment analysis by converting sentences to numerical vectors: a twitter case study," *Multimed. Tools Appl.*, 2019, doi: 10.1007/s11042-019-7586-4.
- [12] J. Awwalu, A. A. Bakar, and M. R. Yaakub, "Hybrid N-gram model using Naïve Bayes for classification of political sentiments on Twitter," *Neural Comput. Appl.*, 2019, doi: 10.1007/s00521-019-04248-z.
- [13] S. R. Ahmad, A. A. Bakar, and M. R. Yaakub, "A review of feature selection techniques in sentiment analysis," *Intell. Data Anal.*, vol. 23, no. 1, pp. 159–189, 2019, doi: 10.3233/IDA-173763.
- [14] A. Unler and A. Murat, "A discrete particle swarm optimization method for feature selection in binary classification problems," *Eur. J. Oper. Res.*, vol. 206, no. 3, pp. 528–539, Nov. 2010, doi: 10.1016/j.ejor.2010.02.032.
- [15] K. Ganesan, "Text Preprocessing for Machine Learning & NLP," Kavita Ganesan - Build Beautiful NLP & Data Applications., 2019. [Online]. Available: <https://kavita-ganesan.com/text-preprocessing-tutorial/#.XrcyMxMzab8>. [Accessed: 12-May-2020].
- [16] K. Akshi and K. Renu, "Firefly Algorithm for Feature Selection in Sentiment Analysis," *Comput. Intell. Data Mining, Adv. Intell. Syst. Comput.*, vol. 556, pp. 693–703, 2017, doi: 10.1007/978-981-10-3874-7.
- [17] M. Tubishat, M. A. M. Abushariah, N. Idris, and I. Aljarah, "Improved whale optimization algorithm for feature selection in Arabic sentiment analysis," *Appl. Intell.*, vol. 49, no. 5, pp. 1688–1707, 2019, doi: 10.1007/s10489-018-1334-8.
- [18] A. Krouska, C. Troussas, and M. Virvou, "The effect of preprocessing techniques on Twitter sentiment analysis," *IISA 2016 - 7th Int. Conf. Information, Intell. Syst. Appl.*, 2016, doi: 10.1109/IISA.2016.7785373.
- [19] W. Shang, H. Huang, H. Zhu, Y. Lin, Y. Qu, and Z. Wang, "A novel feature selection algorithm for text categorization," doi: 10.1016/j.eswa.2006.04.001.
- [20] Z. Zhai, B. Liu, H. Xu, and P. Jia, "Constrained LDA for grouping product features in opinion mining," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2011, vol. 6634 LNAI, no. PART 1, pp. 448–459, doi: 10.1007/978-3-642-20841-6_37.
- [21] A. Bagheri, M. Saracee, and F. De Jong, "Care more about customers: Unsupervised domain-independent aspect detection for sentiment analysis of customer reviews," *Knowledge-Based Syst.*, vol. 52, pp. 201–213, Nov. 2013, doi: 10.1016/j.knosys.2013.08.011.
- [22] Z. A. A. Alyasseri, A. T. Khader, M. A. Al-Betar, M. A. Awadallah, and X. S. Yang, "Variants of the flower pollination algorithm: A review," *Stud. Comput. Intell.*, vol. 744, pp. 91–118, 2018, doi: 10.1007/978-3-319-67669-2_5.
- [23] M. H. Aghdam, N. Ghasem-Aghaee, and M. E. Basiri, "Text feature selection using ant colony optimization," *Expert Syst. Appl.*, vol. 36, no. 3 PART 2, pp. 6843–6853, Apr. 2009, doi: 10.1016/j.eswa.2008.08.022.
- [24] M. H. Aghdam and S. Heidari, "Feature Selection Using Particle Swarm Optimization in Text Categorization," *J. Artif. Intell. Soft Comput. Res.*, vol. 5, no. 4, pp. 231–238, Oct. 2015, doi: 10.1515/jaiscr-2015-0031.
- [25] S. Nemati, M. E. Basiri, N. Ghasem-Aghaee, and M. H. Aghdam, "A novel ACO-GA hybrid algorithm for feature selection in protein function prediction," *Expert Syst. Appl.*, vol. 36, no. 10, pp. 12086–12094, Dec. 2009, doi: 10.1016/j.eswa.2009.04.023.
- [26] Z. Abdi, A. Alyasseri, A. T. Khader, M. A. Al-betar, M. A. Awadallah, and X. Yang, "Nature-Inspired Algorithms and Applied Optimization," vol. 744, no. October 2017, 2018, doi: 10.1007/978-3-319-67669-2.
- [27] J. Too and S. Mirjalili, "A Hyper Learning Binary Dragonfly Algorithm for Feature Selection: A COVID-19 Case Study," *Knowledge-Based Syst.*, vol. 212, p. 106553, Jan. 2021, doi: 10.1016/j.knosys.2020.106553.
- [28] B. Seerat and F. Azam, "Opinion Mining: Issues and Challenges (A survey)," *Int. J. Comput. Appl.*, vol. 49, no. 9, pp. 975–8887, 2012, doi: 10.5120/7658-0762.
- [29] U. Pervaiz, S. Khawaldeh, T. A. Aleef, V. H. Minh, and Y. B. Hagos, "Activity monitoring and meal tracking for cardiac rehabilitation patients," *Int. J. Med. Eng. Inform.*, vol. 10, no. 3, pp. 252–264, 2018, doi: 10.1504/IJMEI.2018.093365.
- [30] S. Asghari and N. J. Navimipour, "Review and Comparison of Meta-Heuristic Algorithms for Service Composition in Cloud Computing," 2015.
- [31] B. O. Aljila, C. P. Lim, L. P. Wong, A. T. Khader, and M. A. Al-Betar, "An ensemble of intelligent water drop algorithm for feature selection optimization problem," *Appl. Soft Comput. J.*, vol. 65, pp. 531–541, Apr. 2018, doi: 10.1016/j.asoc.2018.02.003.
- [32] A. S. Ghareb, A. A. Bakar, and A. R. Hamdan, "Hybrid feature selection based on enhanced genetic algorithm for text categorization," *Expert Syst. Appl.*, vol. 49, pp. 31–47, May 2016, doi: 10.1016/j.eswa.2015.12.004.
- [33] X. Yang, "Flower Pollination Algorithm for Global Optimization," pp. 240–249, 2012.
- [34] P. H. P. Sarijiya and T. A. Saputra, "Modified Flower Pollination Algorithm for Non smooth and Multiple Fuel Options Economic Dispatch," in *2016 8th International Conference on Information Technology and Electrical Engineering (ICITEE)*, 2016, pp. 346–350.
- [35] F. P. Sakti, S. Sarijiya, and S. P. Hadi, "Optimal Power Flow Using Flower Pollination Algorithm: A Case Study of 500 kV Java-Bali Power System," *IJITEE (International J. Inf. Technol. Electr. Eng.)*, vol. 1, no. 2, pp. 45–50, Sep. 2017, doi: 10.22146/ijitee.28363.
- [36] D. Rodrigues, G. F. A. Silva, J. P. Papa, A. N. Marana, and X.-S. Yang, "EEG-based person identification through Binary Flower Pollination Algorithm," *Expert Syst. Appl.*, vol. 62, pp. 81–90, Nov. 2016, doi: 10.1016/j.eswa.2016.06.006.
- [37] Z. A. E. M. Dahi, C. Mezioud, and A. Draa, "On the efficiency of the binary flower pollination algorithm: Application on the antenna positioning problem," *Appl. Soft Comput. J.*, vol. 47, pp. 395–414, Oct. 2016, doi: 10.1016/j.asoc.2016.05.051.
- [38] P. Agarwal and S. Mehta, "Enhanced flower pollination algorithm on data clustering," *Int. J. Comput. Appl.*, vol. 38, no. 2–3, pp. 144–155, 2016, doi: 10.1080/1206212X.2016.1224401.
- [39] R. Douglas, X.-S. Yang, A. N. de Souza, and J. P. Papa, "Binary Flower Pollination Algorithm and Its Application," in *Recent Advances in Swarm Intelligence and Evolutionary Computation*, no. April 2016, 2015, p. 303.
- [40] S. A.-F. Sayed, E. Nabil, and A. Badr, "A binary clonal flower pollination algorithm for feature selection," *Pattern Recognit. Lett.*, vol. 77, pp. 21–27, 2016, doi: 10.1016/j.patrec.2016.03.014.
- [41] H. M. Zawbaa, A. E. Hassanien, E. Emary, W. Yamany, and B. Parv, "Hybrid flower pollination algorithm with rough sets for feature selection," *2015 11th Int. Comput. Eng. Conf. Today Inf. Soc. What's Next?, ICENCO 2015*, pp. 278–283, 2016, doi: 10.1109/ICENCO.2015.7416362.
- [42] H. Majidpour and F. G. Soleimani, "An Improved Flower Pollination Algorithm ...," Sari Branch, Islamic Azad University, Feb. 2018.
- [43] B. Glover, *Understanding Flowers and Flowering: An integrated approach*. Oxford University Press, 2008.
- [44] M. L. Kawasaki and A. D. Bell, "Plant Form. An Illustrated Guide to Flowering Plant Morphology.," *Brittonia*, vol. 43, no. 3, p. 145, Jul. 1991, doi: 10.2307/2807042.
- [45] F. B. Ozsoydan and A. Baykasoglu, "Analysing the effects of various switching probability characteristics in flower pollination algorithm for solving unconstrained function minimization problems," *Neural*

- Comput. Appl., vol. 31, no. 11, pp. 7805–7819, Nov. 2019, doi: 10.1007/s00521-018-3602-2.
- [46] I. Pavlyukevich, “Lévy flights, non-local search and simulated annealing,” *J. Comput. Phys.*, vol. 226, no. 2, pp. 1830–1844, Oct. 2007, doi: 10.1016/j.jcp.2007.06.008.
- [47] X. S. Yang, M. Karamanoglu, and X. He, “Flower pollination algorithm: A novel approach for multiobjective optimization,” *Eng. Optim.*, vol. 46, no. 9, pp. 1222–1237, 2014, doi: 10.1080/0305215X.2013.832237.
- [48] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification and regression trees*. CRC Press, 2017.
- [49] A. Kumar, A. Jaiswal, S. Garg, S. Verma, and S. Kumar, “Sentiment Analysis Using Cuckoo Search for Optimized Feature Selection on Kaggle Tweets,” *Int. J. Inf. Retr. Res.*, vol. 9, no. 1, pp. 1–15, 2018, doi: 10.4018/ijirr.2019010101.

Re-CRUD Code Automation Framework Evaluation using DESMET Feature Analysis

Asyraf Wahi Anuar¹, Nazri Kama², Azri Azmi³, Hazlifah Mohd Rusli⁴, Yazriwati Yahya⁵

Advanced Informatics Department, Razak Faculty of Technology and Informatics

Universiti Teknologi Malaysia, 54100 Kuala Lumpur, Malaysia^{1, 2, 3, 4, 5}

Faculty of Information Management, Universiti Teknologi MARA Cawangan Johor, Kampus Segamat, Johor, Malaysia¹

Abstract—A unified view of web application design and development is crucial for dealing with complexity. However, the literature proposes many denominations, depending on the development methodology, frameworks or tools. This multitude of Create, Read, Update and Delete (CRUD) approaches does not allow a holistic view of the web application. Besides, in a web application, the search for good practice in design, features and essential functions is still a relevant issue. A subset of essential CRUD operations is to provide code automation for web application rapid prototyping. Re-CRUD articulates the records management features into CRUD operation. This study aims to provide insight into the effectiveness and efficiency of Re-CRUD in web application development and to compare it with other web application frameworks' CRUD output. The qualitative feature analysis is used based on the evaluation guideline proposed in DESMET and reviewed by experts for validation. A document management system is developed and used as a case study for Re-CRUD evaluation. The feature analysis comprises Re-CRUD and four other web application frameworks CRUD, namely, CakePHP, Laravel, Symfony and FuelPHP. According to the review, Re-CRUD satisfies its expectations by providing more useful features and delivering higher code automation in the web application development process. Compared to the other existing CRUD generator, Re-CRUD has integrated records management features that are useful in providing support in managing born-digital data and also contributes to effectiveness and efficiency in web application development.

Keywords—Re-CRUD; web application; DESMET feature analysis; electronic records features

I. INTRODUCTION

Web technologies have significantly influenced web application (WA) development and information system. The innovation of web technologies has allowed software developers and engineers to develop responsive cross-platform web applications rapidly. These technologies include creating read update delete (CRUD) generator, web application framework (WAF) and libraries that promotes reusable codes, rapid development and feasible features. WA can be considered a software component that stores and manages information just like a traditional information system but uses explicitly the web paradigm and associated technologies [1], [2]. It is a software system whose primary purpose is to publish and maintain data using hypertext-based principles [3], [4]. WA offers ease of access, maintenance, and cross-platform compatibility compared to the traditional desktop application installed on a local computer [5]. WA can be characterized as one that uses Web architecture and other technologies

(database, browser) to construct an information system that serves organizational needs [6].

The technique of WA development has rapidly changed over these past few years with the born of WAF, which promotes better development experiences and resources management [7], [8]. The WA development has become easier with the adoption of the frameworks, and lots of web-based innovation has been produced by non-technical people due to the framework innovation [9]–[13]. The WAF has bridged the possibility of WA development by a non-programmer and unlocks many potential new software possibilities and ideas [9], [14]. Most of the WAF embraces the MVC architecture that supports rapid and parallel development through CRUD operation, asynchronous technique, and straightforward business logic implementation, making the WA development more practical [9]–[12].

CRUD is the four fundamental components that manage the web application (WA) resources [12], [13], [15]–[18]. The create component allows the user or the WA itself (or both) to add a new data item to the database. Read component is used to retrieve the items recorded in the database and render them into a web page. The update component enables the user or the WA to edit an existing item and have those changes written back to the database. The last component, delete, enable the user or the WA to remove an item record from the database [19].

The CRUD paradigm is widely used among WA developers since it allows them to construct basic WA routine code and define how items in WA are related to one another [20]–[22]. CRUD is a provision of assistance in code generation and basic functionalities to support the developer in accomplishing the task [23]. CRUD enables the developer to create a quick-start application to work as the foundation of the WA solution [7], [24]. CRUD is an excellent technique to start an MVC-based WA project as it provides automation in design patterns [25]. Further, CRUD is a handy time-saver. It generates the skeleton codes for the WA and enables the developer to get faster output and demonstrate the WA prototype (input, process and output) to the WA project stakeholder [26].

The implementation of CRUD in WA development provides a substantial productivity boost for developers [12], [13], [15]–[17]. Using CRUD, the developer does not have to worry about many of the subtle details of wiring up the controller for the MVC application [12], [13], [27]. Although it boosts the development process, the traditional CRUD

generator only generates the fundamental functions that still present problems: their inadequacy to deal with the form features, authentication, search, file management, and others. [12], [13], [28]–[30]. A standalone CRUD could not satisfy the development of decent and complex WA since it lacks standard modern WA features to support the functions such as authentication, authorization, files management, search, internationalization, form features, report, logging and others [12], [13], [29]. Further enhancement and manual code modification are required to improvise the half-baked generated CRUD, especially in the integration of the time-consuming features and comprise repetitive coding for each of the CRUD output [12].

There is also an argument that the CRUD operation is not yet a complete solution for web application development. Many redundant tasks include repetitive code modification for feature integration after generating the CRUD [12]. Coding the same routine code for WA features repeatedly takes a long time and increases development costs [31]. However, it has a huge potential to go beyond the limit where it can automate more components for WA development [32]. The primary purpose of this study is to provide insight into Re-CRUD effectiveness and efficiency in web application development using the qualitative feature analysis case study based on the proposed guide in DESMET [33]. The contributions of this work are as follows:

- This work implements the records management features into CRUD operation for web application development.
- This study analyzes the effectiveness and efficiency of Re-CRUD in providing more features for web application development.
- This work compares the features and output of Re-CRUD and other aristocratic CRUD generators in web application frameworks.

The article is organized as follows. Section 2 describes the main related studies that have been conducted in the field of web application development challenges and Re-CRUD. Section 3 presents the methodology, including the DESMET feature analysis procedure and the instrumentation. Section 4 describes and analyzes the results. Section 5 presents a brief discussion. Finally, Section 6 presents the conclusions.

II. LITERATURE REVIEW

A. Issues and Challenges in Web Application Development

Web application (WA) development is a complex and challenging task since it requires consideration of numerous factors and requirements, some of which may contradict [34]–[42]. Many researchers have widely discussed the scalability issues, and developing a WA that scales well is a challenge [34], [43], [44]. As the WA is becoming mission-critical, there is a greater demand for it to improve reliability, effectiveness, performance, integration and security [34], [45]–[48]. To comply with the diverse expectations and requirements of many different users with varying skills and knowledge is very challenging. Most of the end-users are visual-oriented, which focuses on WA having more multimedia elements instead of

focusing on the functions of WA. They expect to manage and find the information they need faster using WA [6], [34].

Most WA is designed with a WAF with multi plugins and third-party scripting to enhance the features and a database as the storage medium [24], [26], [49]. Integrating plugins requires extensive knowledge of the amalgamation of the plugin and the WA due to the different programming languages and processes [50]–[52]. The integration is crucial as it will affect the WA's performance and stability, including the WA's functionality [34]. The developer must ensure the plugin's compatibility with the WAF and standard web browser technology [49], [53], [54]. Incompatible plugin integration may lead to vulnerabilities and security issues for WA due to deprecated methods or coding, leading to data corruption and unstable WA processing [55]–[58].

Due to the overwhelming number of WAFs published can be a daunting task to determine the most appropriate WAF [59]. The most important is that it can increase the programming productivity through code and files generation automation using CRUD operation, security advantages, and open-source that will impact the cost, support and documentation [60]. The WAF selection is vital as the speed and quality of work depend on it [60], [61].

The usability features issues in WA are design layout, design consistency, accessibility, information content, navigation, personalization, performance, reliability, filtering, analysis and design standards [62]–[66]. User response reveals that the search functionality, consistent navigation throughout the system, authentication, authorization, responsiveness and data visualization and reporting aid most WA usability [34], [67], [68]. Search features have been vital components in WA as the rapid trends in born-digital data and information lead to information overload and exposure [69]. The proper search function enables the user to filter the relevant information based on the search query [69]–[71]. The search function is a complex component where the developer needs to understand the filtering algorithm to ensure that the search results respond to the requested queries [11], [72]. The common issue in the search function is unable to satisfy the user query request due to the incomplete filtering algorithm [34], [62].

Localization is adapting WA (regularly written in English) for use in other countries, considering their culture, standard, regulations, principles and technology conditions [34], [73], [74]. Localization is more than just a language translation, and WA needs to be precisely designed to accomplish this multifaceted condition [73], [75], [76]. The scalability, reliability, availability, maintainability, upgradability, usability, speed and security are the terms used to describe how well the WA meets current and future needs. These iltities describe WA architectural qualities [34], [46], [77]–[79].

The design and development of WA for mobile compatible and device-independent operations are very complex and challenging. It must address various additional aspects compared to the traditional information system or desktop application and needs to satisfy many different stakeholders besides the diverse range of users. Poorly designed and develop WA has a high probability of low performance or failure [80].

B. Re-CRUD

Re-CRUD was generally derived from the absence of electronic records management important features in the WAF CRUD. The conventional WAF CRUD operation is limited to only producing fundamental functions for WA, which requires further modification to include the other useful WA features to ensure the content can be managed systematically. Re-CRUD integrates the CakePHP framework CRUD operation with electronic records management important features that can produce additional useful functions and code automation. The following electronic records management important features are included in Re-CRUD [81]:

- Inventory: a descriptive listing of each record series or system, together with an indication of the location, access, and other pertinent data [82]–[84].
- Retention schedule: list how long each record series must be kept (the retention period), when the retention period starts (the cut-off), and the proper way to dispose of the record once retention is met (the disposition method) [84]–[87].
- Appraisal: the process of determining the archival value and ultimate disposition of records. Appraisal decisions are informed by several factors, including the historical, legal, operational, and financial value of the records [88]–[90].
- Disposition: the process of destruction of records or the transfer of records to another entity (most commonly an archival repository) for permanent preservation [91], [92].
- Role-based access control (RBAC): provides a role-based access control mechanism to offer protection from unauthorized access. Authenticated users with different roles have different authorization or access to the records [93]–[95].
- Search and retrieval: Enables the user to locate and retrieve records based on specific metadata, words or phrases. It is vital in any WA as it enables fast data retrieval via the search parameter [96], [97].
- Audit trail: provides log tracking for any changes to the electronic records to ensure validity and integrity [98], [99].
- Digital Archiving: transfer and store the valuable records into a repository that makes it non-active but still accessible through the system. It also helps to reduce the cluttered old and non-active records from the system [100], [101].
- Sharing: it provides the ability to transfer the record (internal to external or external to internal) in a single data or bulk data. There are several suggested formats such as CSV, XML and JSON [102], [103].
- Reporting: It summarises the current status of records such as total, active, inactive and the required appraisal attention and others [104], [105].

- Others: focused on the front-end framework for UI, data visualizer for reporting page, jQuery, DOMPDPF and others supporting UI features. [82], [106].

Fig. 1 provides a simplified overview of Re-CRUD design based on the CRUD evolution using the console framework. It shows the evolution of the CRUD generation based on the traditional CRUD introduced in 1983 and is mainly used for the database abstraction process [17]. The technology continues to evolve into other application development segments, such as web application CRUD.

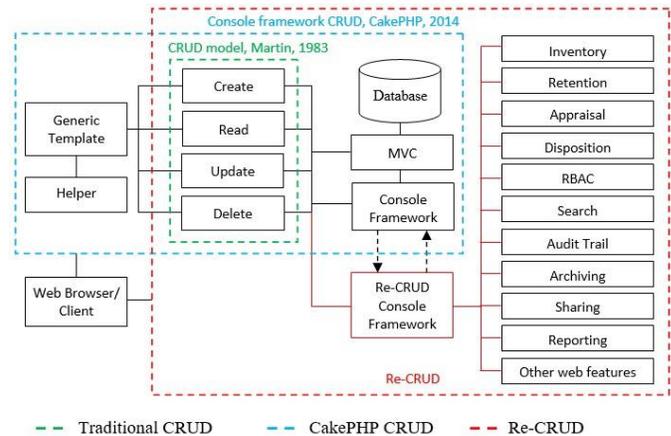


Fig. 1. Re-CRUD Console Framework.

III. METHODOLOGY

A. DESMET Feature Analysis

The DESMET method is designed to assist an evaluator in preparing and carrying out an unbiased and reliable assessment exercise within the framework of software engineering approaches and tools [33]. Feature Analysis is the term used by DESMET to describe a qualitative evaluation. Feature analysis focused on recognizing the requirements for a particular task and mapping it to the tool that was used to solve the task. Feature analysis is qualitative because it requires a subjective assessment of the relative importance of different features and how well they have been implemented. The qualitative case study is a feature-based evaluation performed by someone who has used the tool on an actual project. There are three main processes involved in carrying out a feature analysis, i) selection of feature; ii) feature scoring and ranking; iii) analysis and result interpretation, as explained in section V. For this research, the feature analysis is executed by the i) author and; ii) selected respondents, to compare and validate the feature analysis scoring from author bias. The following sub-section discusses the DESMET feature analysis procedure.

1) *Selection of features:* For features selection, the Re-CRUD features are grouped into four categories: CRUD operation, RBAC, electronic records management, and others, as shown in Table I. The sub-features are categorized based on the domain of the function.

2) *Feature scoring and ranking:* The Judgement Scale and Interpretation (JIS) is another definition that must be completed using DESMET. As described by Kitchenham [33], the JIS evaluation is according to the evaluator's perception.

As shown in Table II, if the feature is fully supported, the JIS is 1; if the feature is partly supported, JIS is 0.5; if the feature is not supported, JIS is 0 and -1 if the features make things worst [33]. The importance of a feature can be determined by deciding whether it is mandatory or merely desirable.

A method or tool that lacks a mandatory feature is unacceptable [33], [107], [108]. Non-mandatory features are considered desirable. This viewpoint on importance leads to two evaluation criteria: one that decides whether or not a feature is mandatory and the other that determines the degree to which a non-mandatory feature is desired. A ranking method is used to identify the electronic record important features are vital or only a desire for WA. Table III present the features set important weightage to identify the most imperative integrated features.

TABLE I. FEATURE SELECTION

Features		Sub-Features		Description
F1	CRUD operation	SF01	Code automation	The automation of code generation for persistent storage
F2	RBAC	SF01	RBAC	Provides a role-based access control mechanism to offer web content protection from unauthorized access.
F3	Electronic Record Mgt	SF01	Inventory	a descriptive listing of each record series, together with an indication of the location, access and other pertinent data
		SF02	Search	Ability to locate and retrieve data/record based on specific metadata, keyword or phrases.
		SF03	Audit Trail	provides log tracking for any changes to the electronic records to ensure that validity and integrity
		SF04	Transfer & Sharing	enable transfer or share the data (internal to external or external to internal) in a single or bulk data.
		SF05	Report	provides a summary related to the current status of records, active or inactive.
		SF06	Retention	list how long each record series must be kept
		SF07	Appraisal	the process of determining the archival value and ultimate disposition
		SF08	Archive	the records will be permanently stored, inactive but accessible for future references
		SF09	Disposition	the process of permanent destruction of records.
F4	Other web features	SF01	UI, visual design	the UI and other supporting components that supports the web application design and data presentation

TABLE II. JUDGMENT SCALE & INTERPRETATION (JI SCORE) [33]

Generic scale point	Definition of scale point	Score
Make things worse	Cause Confusion. The way the feature is implemented makes it difficult to use and/or encourages incorrect use.	-1
No support	Fails to recognize it. The feature is not supported nor referred to in the user manual.	0
Partly support	The feature is supported indirectly, for example, by using other tool features in non-standard combinations.	0.5
Full support	The feature appears explicitly in the feature list of the tools and user manual. All aspects of the feature are covered.	1

TABLE III. FEATURE SET IMPORTANT WEIGHTAGE [33]

Acronym	Level of importance	Weightage
M	Mandatory	4
HD	Highly desirable	3
D	Desirable	2
N	Nice to have	1

3) *The feature set and total score:* DESMET is an assessment technique that requires the assignment of scores to features and sub-features such as sub-feature importance levels, feature weights, and judgement scales. According to Marshall [108], the author first determined each score. At the initial stage, DESMET is required to determine the important level for each sub-feature. The identified features and sub-features set important weighting used in the feature analysis is shown in Table IV.

TABLE IV. FEATURES AND SUB-FEATURES SCORE USED IN THE ANALYSIS

ID	Feature Set	SF-ID	Subfeature: level of importance	Feature set Importance Weighting
F1	CRUD operation	F1-SF01	Mandatory	4
F2	RBAC	F2-SF01	Highly Desirable	3
F3	Electronic Record Management	F3-SF01	Desirable	2
		F3-SF02	Mandatory	4
		F3-SF03	Mandatory	4
		F3-SF04	Desirable	2
		F3-SF05	Highly Desirable	3
		F3-SF06	Highly Desirable	3
		F3-SF07	Mandatory	4
		F3-SF08	Mandatory	4
		F3-SF09	Mandatory	4
F4	Other web features	F4-SF01	Highly Desirable	3

The level of importance, together with the weightage for each of the sub-features, is identified. A feature will receive the highest possible score if all of the features in the set are completely present or supported. The level of importance for the features is determined by their implication and significance [108]. These weighted scores can be summed to obtain a percentage (%) score for each list of features. The following equation is used to compute the rating score for a feature set:

$$future\ set\ score\ \% = \frac{\sum\ of\ feature\ set\ score}{maximum\ score} \times 100 \quad (1)$$

In Table IV, the feature set is divided into four, F1 comprise one sub-features (F1-SF01), F2 comprises one sub-features (F2-SF01), F3 comprises nine sub-features (F3-SF01 to F3-SF09), and the F4 comprises one sub-feature (F4-SF01). The average (weighted) rate scores for each feature set are used to calculate a general percentage score for each model. A normalized score (percentage) is utilized because the feature set's sub-features vary.

For this calculation, the feature set weighting is used in Table V. The values here emphasize support for code automation via CRUD operation (F1) and for the electronic records management features (F3). Other weightings could be used, perhaps to emphasize usability, as tools to support the proposed solution to become more mature. The overall score for each CRUD generator can be determined using the following equation:

$$Overall\ score = \frac{\sum_{i=1}^4 (W_i TP_i)}{\sum_{i=1}^4 (W_i)} \quad (2)$$

W_i is the weighting for the i^{th} feature set and TP_i is the percentage (%) score for the i^{th} feature set.

TABLE V. FEATURE SET WEIGHTING

Feature Set	Weight
F1	0.4
F2	0.2
F3	0.3
F4	0.1

B. Instrumentation

The primary purpose of the evaluation is to provide insight into Re-CRUD effectiveness and efficiency. A qualitative feature analysis case study is used based on the evaluation of guidelines proposed in DESMET [33] and reviewed by experts for validation.

1) *Case study:* Electronic Document Management System (EDMS) is used as a case study to assess and validate Re-CRUD. EDMS is a software system used to manage (organize and store) different kinds of data, information, and records. For this research, the EDMS is focused on managing born-digital data where it should be able to:

- Capture and validate the data input.
- Protect content using authentication and authorization.

- Practice proper electronic records management for the content.
- Multi-device and platform friendly.

Fig. 2 shows the application module that is available in the EDMS. To ensure the content is protected, the user must register and authenticate before accessing the content. It also includes the authorization procedure. Most of the records management aspect is put into practiced in the document repository.

As discussed in the following sub-section, seven steps are involved in designing and administering the case study [52].

a) *Identify case study context:* The case study objective is to evaluate the effectiveness and efficiency of the proposed solution. Table VI [55]–[58] shows that effectiveness and efficiency are highlighted as usability features. In the context of WA development, Re-CRUD is used to generate codes and files to form a WA with integrated records management features, as highlighted in Table IV.

b) *Select the host projects:* The EDMS case study applies the potential electronic record important aspects and is used to evaluate its effectiveness and efficiency [109].

c) *Identify the method of comparison:* As a comparative method, a cross-platform comparison is performed. Four additional CRUD generators are used for the cross-platform comparison, one using Re-CRUD and the others using the existing CRUD generator. The characteristics of all other development methodologies and procedures will be the same.

d) *Minimize the effect of confounding factors:* A similar host project characteristic and data storage are adopted to minimize the confounding effects. The selected WAF CRUD also have the same programming language, shares the same web server environment and uses almost similar development architecture. The selected respondents also must have experience, knowledge and understanding of software development and the use of WAF together with the CRUD operation [110]. This ensures that they are familiar with the CRUD operation and can focus on the evaluation instead of learning how to perform the CRUD operation and reduce the learning curve. Besides, specific and detailed instruction on configuration, database schema, development method, and CLI command is provided for the case study development.

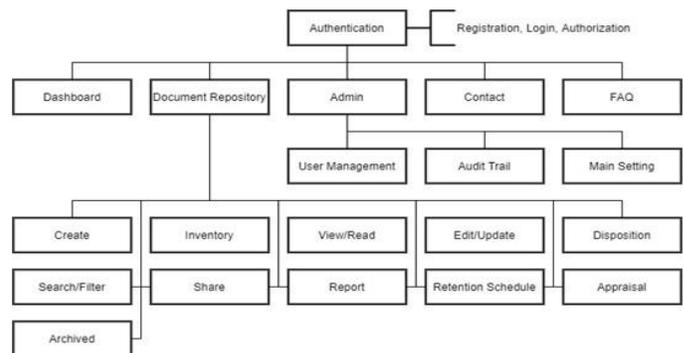


Fig. 2. EDMS Application Module used in Feature Analysis Case Study.

TABLE VI. SYSTEM USABILITY SCALE (SUS) QUESTIONNAIRE [115]

Usability aspect		Candidate item
Effectiveness (ES)	ES1	It allows me to accomplish my tasks. I think I would not need a system with more features for my tasks. I would not need to supplement Re-CRUD with an additional components. I found the system unnecessarily complex This system's capabilities would meet my requirements.
	ES2	
	ES3	
	ES4	
	ES5	
Efficiency (EC)	EC1	It saves me time when I use it. I found the various functions in this system were well integrated. I tend to reduce a lot of mistakes with this system. I don't make many errors with this system. I don't have to spend a lot of time correcting things with this system
	EC2	
	EC3	
	EC4	
	EC5	

e) *Plan the case study*: The following activities must be sequentially performed to complete the EDMS case study development. The listed activities require an understanding of software installation, configuration, MySQL database and PHP programming language.

- Gathering required files from the Github repository
- Host configuration
- Database migration and seeding
- Performed CRUD operation
- Output evaluation

f) *Executing the case study*: The author and respondents execute the development of the EDMS based on the given instruction and software development specifications. The respondents will fill in the online evaluation form at the end of the process.

g) *Analyze and report the results*: The score in the evaluation will be consolidated to identify the effectiveness and efficiency of the integrated electronic records management features in WA development.

2) *Expert validation*: Expert validation comprises opinion and judgement from the individual with knowledge and experience in the subject matter [111]–[113]. Expert validation is a methodology in which judgment is based on a particular set of requirements and/or experience obtained in a specific knowledge field, application area, product area, specific discipline, sector, and others [114]. The focus of this process is to validate the usability of the proposed solution. To execute the expert validation for this research, the experts must possess knowledge in software development, testing, maintenance and web-based technologies. The experts comprise representation from the industry, public sector and academician. The selected expert must respond to a set of questions as discussed in the following section.

3) *System usability scale questionnaire*: A set of questionnaires is provided to the respondents which are designed based on the System Usability Scale (SUS)

questionnaire [115]–[117]. The System Usability Scale (SUS) is an inexpensive yet effective and reliable tool for assessing the effectiveness and efficiency of a product [115]–[121].

The questionnaire is divided into two sections. First, the demographic comprises questions on the highest qualification, current working position, software development experiences, sector (mobile, web, IoT, desktop, cloud) and primary programming language. The second section will include the effectiveness and usability instrument and open-ended comments. The rating is based on 5 points Likert scale, which is anchored with one as strongly disagree and five as strongly agree [118], [122]–[124]. Table VI shows the usability aspect and the candidate item mentioned in the SUS questionnaire.

The SUS result is interpreted based on the grading scale, as shown in Fig. 3, to get a clearer picture of the effectiveness and efficiency of the proposed solution.

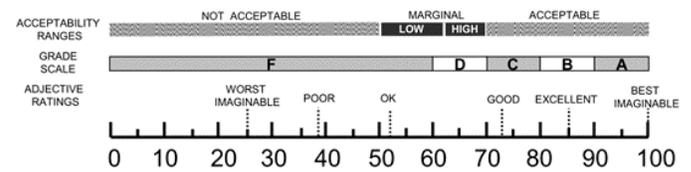


Fig. 3. The Adjective Ratings, Acceptability Scores and Grading Scales in Relation to the Average SUS Score [119].

The following equation is used to calculate the SUS score. The result will be interpreted based on Fig. 3.

$$sus\ score = \frac{(x \times 1) + (x \times 2) + (x \times 3) + (x \times 4) + (x \times 5)}{\text{total respondent} \times 5} \times 100 \quad (3)$$

IV. ANALYSIS AND RESULT

This section summarises the total scores for each CRUD generator for web application development. Table IV shows the feature weighting, and Table VII summarises the results of feature analysis for each of the WAF CRUD generators. This study aims to assess the effectiveness and efficiency of integrated records management aspects in CRUD operation from a web application development perspective. The feature analysis comprises five CRUD generators (embedded in WAF), i) Re-CRUD; ii) CakePHP; iii) Laravel; iv) Symfony and; v) FuelPHP. At the initial stage of DESMET evaluation, the importance level score has been given for each feature set as mentioned in Table IV and displayed in the sub-feature weightage score (C). The Judgment Scale and its Interpretation (JIS) is another definition that must be completed using DESMET. As described by Kitchenham [33], the JIS evaluation is according to the evaluator's perception.

Almost all sub-features in Re-CRUD have JIS = 1 since it is designed and developed based on the identified electronic records features important aspects, considering all intended features have been implemented. If the feature is fully supported, JIS is 1. If the feature is partly supported, JIS is 0.5. If the feature is not supported, JIS is 0 and -1 if the features make things worst [33]. Referring to Table VII, columns F, J, N, R and V show the JIS score for each sub-features of Re-CRUD, CakePHP, Laravel, Symfony and FuelPHP, respectively.

TABLE VII. FEATURE ANALYSIS RESULT

					Re-CRUD				CakePHP CRUD				Laravel CRUD				Symfony CRUD				FuelPHP CRUD														
A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y											
Fea-Set	Sub-Fea	SF WS	MF SS	FI W	JI S	SF WS	F S S	FS S%	JI S	SF WS	F S S	FS S%	JI S	SF WS	F S S	FS S%	JI S	SF WS	F S S	FS S%	JI S	SF WS	F S S	FS S%											
F1	F1-SF-01	4	4	0.4	1	4	4	100	0.5	2	2	50	0.5	2	2	50	0.5	2	2	50	0.5	2	2	50											
F2	F2-SF-01	3	3	0.2	0.5	1.5	1.5	50	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0											
F3	F3-SF-01	2	30	0.3	1	2	28	93.33	1	2	9	30	1	2	8	26.67	1	2	9	30	0.5	1	5	16.67											
	F3-SF-02	4			1	4			0.5	2			0.5	2			0.5	2			0.5	2													
	F3-SF-03	4			0.5	2			0	0			0	0			0	0			0	0													
	F3-SF-04	2			1	2			0.5	1			0	0			0	0			0.5	1													
	F3-SF-05	3			1	3			0	0			9	30			0	0			8	26.67			0	0	9	30	0	0	0	0	5	16.67	
	F3-SF-06	3			1	3			0	0			0	0			0	0			0	0			0	0	0	0	0	0	0	0	0	0	0
	F3-SF-07	4			1	4			0	0			0	0			0	0			0	0			0	0	0	0	0	0	0	0	0	0	0
	F3-SF-08	4			1	4			0	0			0	0			0	0			0	0			0	0	0	0	0	0	0	0	0	0	0
	F3-SF-09	4			1	4			1	4			1	4			1	4			1	4			1	4	1	4	1	4	1	4	1	4	1
F4	F4-SF-01	3	3	0.1	1	3	3	100	0.5	1.5	1.5	50	0.5	1.5	1.5	50	0.5	1.5	1.5	50	0.5	1.5	1.5	50											
Total Score					36.5/40				12.5/40				11.5/40				12.5/40				8.5/40														
Feature Set Weighting Overall Score (%)					88.00				34.00				33.00				34.00				30.00														

SFWS - Sub-Feature Weightage Score (Mandatory - 4, Highly Desirable - 3, Desirable - 2, Nice to have - 1) [JIS*SFWS(C)]
 MFSS - Max-Feature Set Score [sum SFWS(C)]
 FIW - Feature Important Weightage (0.4, 0.3, 0.2, 0.1)
 JIS - Judgement Scale and Interpretation (Full supported - 1, Partly supported - 0.5, No support - 0, Make thing worst - -1)
 FSS - Feature Set Score (sum of each SFWS)
 FSS% - Feature Set Score Percentage (FSS/MFSS*100)
 Feature Set Weighting Overall Score - (F1 FSS%*F1 FIW)+(F2 FSS%*F2 FIW)+(F3 FSS%*F3 FIW)+(F4 FSS%*F4 FIW)

Based on the JIS score, the Sub-Features Weightage Score (SFWS) for each WAF CRUD is calculated (Re-CRUD (G); CakePHP (K); Laravel (O); Symfony (S) and; FuelPHP (W)).

The weight score is the multiplication of the JI with the respective Sub-Feature Weightage Score (SFWS) (C). For example, for F1-SF01, the SFWS for Re-CRUD is 4 (G), obtained by JIS(F)*SFWS(C). Next, from the SFWS, the

Feature Set Score (FSS) is calculated (Re-CRUD (H); CakePHP (L); Laravel (P); Symfony (T) and; FuelPHP (X)). This FSS is the sum of the SFWS grouped by the feature set. For example, in column H, the FSS for Re-CRUD feature set 3 (F3) is 28 is the sum of (2+4+2+2+3+3+4+4+4). In the sequence, the FSS percentage (Re-CRUD (I); CakePHP (M); Laravel (Q); Symfony (U) and; FuelPHP (Y)) is obtained by dividing the respective FSS with MFSS (D) for each feature set.

Finally, the overall score for each WAF CRUD is calculated with the sum of each FSS percentage, considering the feature set and sub-feature set important levels (Table VII). For instance, considering the FIW for F1 is 0.4, and the FSS% for F1 Re-CRUD is 100% (I), the % of the feature set weighting score is $100\% * 0.4 = 40\%$. For F2, the weighting score is 0.2; then, it is calculated using $50\% * 0.2 = 10\%$, and so on. In the end, all these values are summed, and the feature set weighting overall score is obtained. Each feature set score will be described in-depth in the following subsection.

Based on the feature set weighting overall score in Table VII, Re-CRUD leads the scores with a massive margin of difference, Re-CRUD: 88%; CakePHP: 34%; Laravel: 33%; Symfony: 34% and; FuelPHP: 30%. Technically, this is because Re-CRUD is specifically designed and developed based on the identified electronic records features mentioned earlier. Rather than integrating the electronic records features into the CRUD operation, Re-CRUD also has reconstructed the existing CRUD operation functions and features to make it more systematic and ensure that each of the features and functions still exists in the code automation generation. The score is within 30% to 34% for the other CRUD generators because the electronic records features are not present in the CRUD operation. Even though some of the features are present, the JIS score is 0.5 (partly supported), requiring modification or enhancement to the generated files and coding. Hence, even though Re-CRUD is new compared to the other listed aristocratic WAF CRUD elite, the integrated electronic records features make it more effective and efficient in WA development, especially in managing the born-digital content in the WA. The integrated important aspects have crucial roles in managing digital content by promoting appropriate electronic records management functions. It enables the semi-active record to be appraised, archived the inactive record, disposed of unused records, reporting, and other features contributing to WA usability.

A. Feature Set 1: CRUD

The CRUD is focused on code automation, where it generates the fundamental function of WA based on the WAF architecture. The CRUD operation is embedded into the respective WAF as a plugin to enable the developer to generate the WA prototype rapidly. During the case study, it was found that all of the listed CRUD generators can produce WA fundamental components as expected. Referring to Fig. 4, the Re-CRUD FSS is 100%, and the other CRUD generators score 50%. Based on the evaluation, all of the CRUD generators can produce a skeleton of WA with fundamental functions. The JIS score for Re-CRUD is 1 (fully supported) compared to the other's score is 0.5 (partly supported). The Re-CRUD operation provides a more comprehensive solution in generating the files

and codes for WA where the additional features from electronic records features are fully integrated and generated using the same command without any additional modification made by the developer. The other WAF CRUD operation is limited to producing the fundamental components only and requires further modification. This modification is considered a manual code modification where the developer must reconstruct some of the CRUD generated code to enhance and integrate with other features or other third-party plugins.

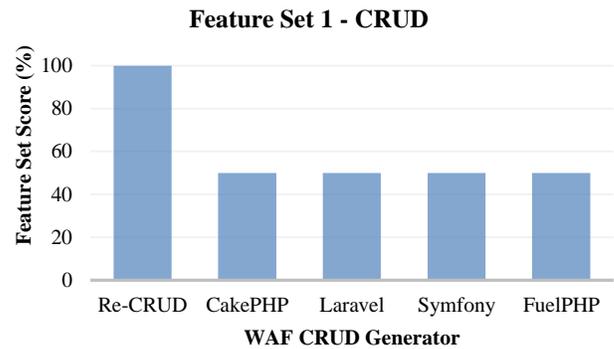


Fig. 4. Feature Set 1 - CRUD Operation (Code Automation) FSS.

From the developer's perspective, the electronic records feature integration is beneficial due to no code modification, reducing the repeating coding process, e.g., integrating search into different tables in WA and reducing the development timeframe. In accordance with the rapid application development methodology in CRUD operation and modern WA features requirements, the integrated electronic records features in Re-CRUD offer an effective solution for rapid development, fully-featured functions and code automation. The additional features generated in CRUD operation allow the developer to focus on the other vital features that the stakeholder requires.

Overall, based on the case study evaluation, all CRUD operations are working as expected and fully supported (code automation) to generate the fundamental component to form a WA. However, Re-CRUD can produce more functions to support the modern WA without having any additional extension or plugin. The MFSS for CRUD is 4 with FIW is 0.4, which considered that the CRUD operation is vital since it is the generator for code automation and crucial to providing a rapid WA development process.

B. Feature Set 2: Role-based Access Control (RBAC)

The RBAC is the authentication and authorization which enable the protection of WA data and content. The features are evaluated to determine whether the respective WAF CRUD can provide the functions without having manual code modification or a third-party plugin. The RBAC is categorized as highly desirable since it is considered one of the most important features that need to be available in WA. Fig. 5 shows the RBAC FSS where the Re-CRUD score is 50%, and the other WAF CRUD is 0%. The RBAC does not fully support the other WAF CRUD generator aptitude. Re-CRUD has embedded the RBAC into the CRUD operation where the process of the authentication data table is migrated and the seed

inside the same generator. However, the integrated RBAC in Re-CRUD is considered partly supported due to the RBAC environment is not entirely present where the authorization policy (who is allowed to access what) is not integrated with CRUD operation. The authorization policy is important to manage and process the user's permission to access a specific resource or function.

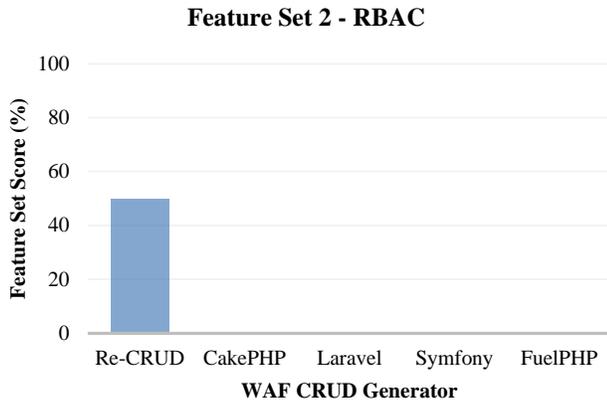


Fig. 5. Feature Set 2 – Role-based Access Control FSS.

The policy needs to be manually generated using the authorization policy. By default, the policy is available, but the CRUD operation did not update it. The developer is expected to have an option to generate the policy content during the CRUD process. This policy automation can be considered in the Re-CRUD future update.

C. Feature 3: Electronic Records Management

Feature set 3 (F3) comprises 9 sub-features, which is, inventory (F3-SF-01); search (F3-SF-02); audit trail (F3-SF-03); transfer and sharing (F3-SF-04); reporting (F3-SF-05); retention (F3-SF-06); appraisal (F3-SF-07); archive (F3-SF-08) and; disposition (F3-SF-09). Fig. 6 shows the highest FSS is by Re-CRUD with a score of 93.33% (Table VII).

Re-CRUD (F3) FSS is 26.5/30, showing that Re-CRUD supports most electronic record features. Based on the JIS, it shows that all of the Re-CRUD F3 sub-features scores are fully supported (JIS = 1) except for F3-SF-03 (audit trail). It is due to the audit trail requiring more in-depth functions to support the audit log process, which enables the edited records to be reverted to the original state and it is considered partly supported (JIS = 0.5). For the CakePHP and Symfony CRUD operation, they share the same FSS score, 30 %, which is 3.33% higher than Laravel's 26.67% score. Compared to the Laravel FSS, the difference between CakePHP and Symfony is the transfer and sharing features where this SFWS is tagged as desirable (2). Both of the WAF CRUD operations partly support (0.5) the feature. FuelPHP FSS for F3 is 16.67%, where most electronic records features are not present. The inventory JIS for FuelPHP CRUD is partly supported (0.5), and for disposition, the JIS is fully supported (1), which brings the SFWS to 5/40. Fig. 7 shows the complete analysis for features 3 (electronic records) components.

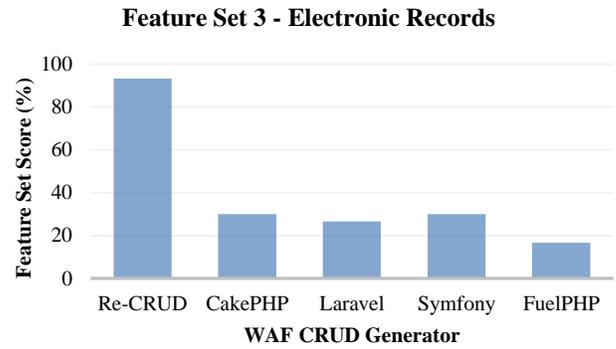


Fig. 6. Feature Set 3 – Electronic Records FSS.

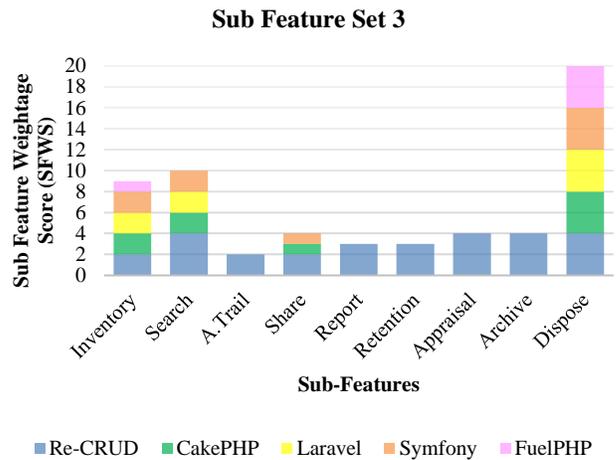


Fig. 7. Feature 3 SFWS (Electronic Records).

TABLE VIII. RESPONDENTS DEMOGRAPHIC DATA

Respondent	Age	Gender	Highest Academic Qualification	Software Development Experience	Software Application Major Sector	Primary Programming Language
R1	36-40	Female	Degree	> 6 years	Web	Java
R2	26-30	Male	Degree	3 - 6 years	Web	PHP
R3	36-40	Female	Master	> 6 years	Web	PHP
R4	36-40	Male	Degree	> 6 years	Web	Python
R5	41-45	Male	Degree	> 6 years	Web	Java
R6	31-35	Male	Degree	> 6 years	Cloud	PHP
R7	36-40	Male	Degree	> 6 years	Cloud	Java
R8	31-35	Male	Master	3 - 6 years	Web	PHP

The audit trails should have a comprehensive function in managing the data history or content changes. The Re-CRUD audit trail aims to preserve the content by tracking the series of changes in the data or records. By tracking the changes, the authenticity of the data or content can be preserved and enable the system administrator to identify the authorized person that amended the content.

The inventory (F2-SF-01) shows that most of the CRUD generator is fully supported (JIS = 1). Most of the CRUD generator produces the index file where this index lists all data or records in the web application, which serve the same purpose of inventory but without the record series. The search (F3-SF-02) shows that Re-CRUD JIS = 1, CakePHP, Laravel and Symfony JIS = 0.5 and FuelPHP JIS = 0. The search feature should be generated in the CRUD operation as the search feature is vital in providing fast access and retrieval. With the integration, Re-CRUD can generate the search function embedded in the inventory and all necessary functions. This is very useful for the developer where they can save more time and reduce the process of coding the search routine code in every single table in the WA. Search is not integrated with the operation for the other web CRUD generator but can be integrated using a plugin. However, the integration required more time and testing the integration.

The sharing feature (F3-SF-04) shows that Re-CRUD has fully supported (JIS = 1), CakePHP and Symfony are partly supported (JIS = 0.5), where the Laravel and FuelPHP JIS = 0 due to not being supported by regular CRUD operation. The sharing feature enables the user to share the content using a link, email or QR code. Re-CRUD generates the sharing function and embeds it into the inventory and view layer. The link and QR code is dynamically generated for each of the content in the web application. However, the access is still subjected to the RBAC to ensure that the content is protected and secure. Without the integrated sharing feature, the developer needs to code the functions for each of the tables, and it requires more time for the development. The reporting feature (F3-SF-05) is considered one of the most crucial features in a web application where it provides a summary of data and records. This feature is useful for the system administrator to populate the data in the web application and retrieve the report, for example, the monthly statistic of the data entry, active and inactive records, archive records and others. Re-CRUD has fully supported this feature (JIS = 1), where the reporting functions are integrated into each generated table. The integration also utilized the data visualization feature from F4-SF-01, where the ChartJS is used to generate the chart to make the report more systematic and readable. However, the reporting features are not supported by the other CRUD generator.

For the retention (F3-SF-06), appraisal (F3-SF-07) and archived (F3-SF-08), it was found that only Re-CRUD is fully supported (JIS = 1), and the other generator is not supported. Retention enables systematic record management in the web application. It provides a duration for each of the records before deciding to be disposed of or archived permanently. The retention feature works together with the appraisal feature (F3-SF-07). The appraisal feature provides the function to evaluate the specific record that has past the retention due date before deciding to be disposed of or archived. Each of these functions is generated through the CRUD operation using Re-CRUD. The digital archival feature in Re-CRUD enables the system administrator to move the inactive record that still has significant value, e.g. fiscal, legal, historical and other vital records, to permanent storage. The records that have moved to the archive are permanently stored, and the edit feature is

disabled to protect the originality and authenticity of the records. This feature is critical to ensure that the web application is not burdened by the unnecessary records, which may lead to an information explosion due to unmanageable data and records in the web application.

The disposition feature (F3-SF-09) shows that all CRUD generators have the feature. The objective of this feature is to remove or dispose of the records from the web application database. Technically, this feature is considered a standard feature in the CRUD operation. Based on the objective, all CRUD operations are fully supported (JIS = 1) since they can delete the records from the web application database.

D. Feature Set 4: Other

Feature set 4 focused on the UI and other supporting features to support the WA design and data presentation. Fig. 8 shows the FSS for each of the WAF CRUD. The weightage score for F4-SF01 is highly desirable since it is important to render the responsive UI, generate charts for data reporting, generate PDF, and provide a WYSIWYG editor. Technically, Re-CRUD is fully supported (100%) with those features and successfully integrated into the CRUD output. For the other WAF CRUD generator, the FSS is 50% since some of the features are available. Although it is not fully similar to the specification, it is still natively able to provide equivalent functions; for example, the template can still render responsively. PDF can be printed using the default print method and others.

E. Expert Validation

To validate the result from Table VII, feature analysis has been carried out with industry professionals' participation to evaluate Re-CRUD. The result is compared with the author's feature analysis result. Expert validation comprises opinion and judgement from the individual with knowledge and experience in the subject matter [111]–[113]. Expert validation is a methodology in which judgment is based on a particular set of requirements and/or experience obtained in a specific knowledge field, application area, product area, specific discipline, sector, and others [114]. The focus of this process is to validate the usability of the proposed solution. To execute the expert validation for this research, the experts must possess knowledge in software development, testing, maintenance and web-based technologies. The experts comprise representation from the industry, public sector and academician. The selected expert must respond to a set of questions as discussed in the following section.

To get a concrete outcome for the case study, it is important to know that the selected respondent must have knowledge in the specific area of the testing and understand how the product can provide a decent solution for their problem [125], [126]. Previous studies have suggested that five respondents are sufficient for usability testing, revealing 80% of usability issues [127]. Another usability researcher stated that a group size of 10 respondents is sufficient where it reveals a minimum of 82% of the problem [128]. A group size of 7 respondents is optimal for studies; even where the study is quite complex will reveal 95% of the problem [129]. It is also suggested that the respondents have expertise in a specific selected field, and the

optimal sample size also should be influenced by the study's complexity and diversity of the respondents [126], [128].

Based on the optimal number of respondents as suggested by previous studies, the usability test for this study is performed by seven selected respondents. The selected respondent is a practitioner that is active and knowledgeable in WA development, where they must-have experience in WA development using WAF and CRUD operation. Considering their experiences and knowledge, the respondents are experts in WA development and familiar with the task domain, WAF, MVC architecture, CRUD and other related features and functions.

The expert validation is executed to compare and validate the feature analysis scoring from the author's bias. 10 Re-CRUD feature analysis case study invitation has been emailed to selected experts, and eight have completed the case study and evaluation. The same case study has been executed by eight selected experts using the same procedure and analyzed. Table VIII shows the respondent demographic data. The demographic information shows that six respondents are male and two female. Their age is in the range of 26 to 45 years old. Six respondents have a bachelor's degree, and the other two respondents have a master's degree. In terms of software development experience, six of the respondents have more than six years of experience, while the others have 3 to 6 years of software development experience, and their software development major sector is in web and cloud. Four of the respondents used PHP as their primary programming language, three respondents used Java, and the others used Python.

Table VII shows the Re-CRUD feature set overall weighting score is 88%, based on the author's evaluation. Fig. 9 compares the Re-CRUD feature set weighting score given by the author and another eight selected experts. A based-line score (88%) has been set and marked with a dashed line compared to the other experts' scores.

Fig. 10 shows the feature analysis responses from industry professionals. Based on the feature set weighting overall score, it was found that 7 (87.5%) of the responses from the experts' score were higher compared to the based-line score (88%) and only 1 (12.5%) scored 0.5 below the based line. Most of the response is above the baseline as the listed features can be generated during the CRUD operation and performed as expected.

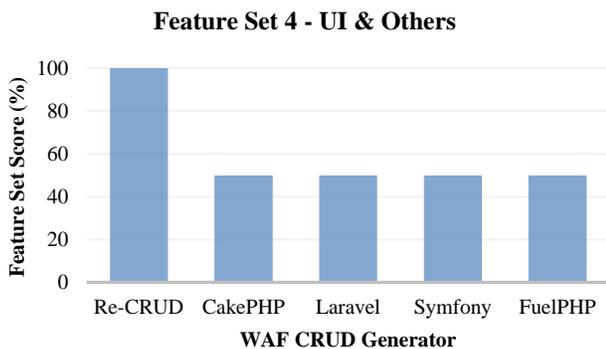


Fig. 8. Feature Set 4 – UI & Others FSS.

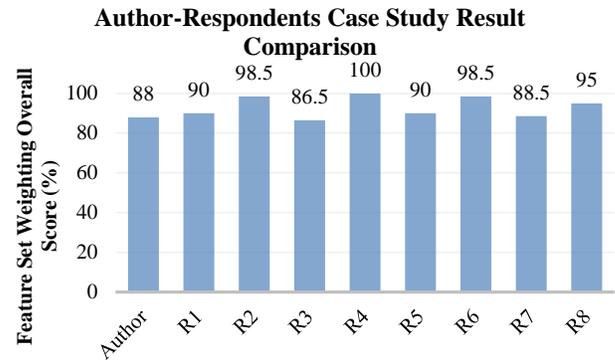


Fig. 9. Comparison of Re-CRUD Feature Set Weighting Score between Author and Respondents.

FS	SF	R1					R2					R3					R4				
		JIS	SFWS	FSS	FSS%	FIW%	JIS	SFWS	FSS	FSS%	FIW%	JIS	SFWS	FSS	FSS%	FIW%	JIS	SFWS	FSS	FSS%	FIW%
F1	F1-SF-01	1	4	4	100	40	1	4	4	100	40	1	4	4	100	40	1	4	4	100	40
F2	F2-SF-01	0.5	1.5	1.5	50	10	1	3	3	100	20	0.5	1.5	1.5	50	10	1	3	3	100	20
F3	F3-SF-01	1	2				1	2				1	2				1	2			
	F3-SF-02	1	4				1	4				1	4				1	4			
	F3-SF-03	1	4				1	4				0.5	2				1	4			
	F3-SF-04	1	2				1	2				1	2				1	2			
	F3-SF-05	1	3	30	100	30	1	3	28.5	95	28.5	0.5	1.5	26.5	88.33	26.5	1	3	30	100	30
	F3-SF-06	1	3				0.5	1.5				1	3				1	3			
	F3-SF-07	1	4				1	4				1	4				1	4			
	F3-SF-08	1	4				1	4				1	4				1	4			
	F3-SF-09	1	4				1	4				1	4				1	4			
F4	F4-SF-01	1	3	3	100	10	1	3	3	100	10	1	3	3	100	10	1	3	3	100	10
FSW Overall Score (%)		90					98.5					86.5					100				

Fig. 10. Feature Analysis Responses.

As highlighted in Fig. 10, several features in the feature analysis score are partly supported. For the F2-SF-01 (RBAC), four respondents (R1, R3, R5 and R7) stated the score was partly supported and tallied with the author has given scores. It was found that the F2-SF-01 (RBAC) feature is considered incomplete because some of the sub-features are not present. As stated in B, the RBAC is partly supported because the authorization policy is not integrated and generated with the CRUD operation. The RBAC policy enables the developer to manage who can access/restrict specific resources. However, four respondents are satisfied with the RBAC, which is fully supported since the authentication is fully functional and the authorization policy is not a compulsory component. Technically, the authorization policy can vary depending on the WA requirement. The authorization can be achieved using the authentication group and session through a simple programming procedure.

One response (R3) stated that F3-SF03 (audit trail) is partly supported. This is due to the issues of the ability to restore the original content from the audit trail history. Re-CRUD is built to have audit trail features that can provide digital tracking of content changes and present a list of changes history. With the

more complex concept of audit trail, R3 expect to have a restoration procedure that can revert specific changes to the original state. This feature is currently not present in Re-CRUD since it was designed to capture and provide the history of the changes. Although the restoration ability can be integrated, it requires a more complex audit trail management. For example, the restoration also needs to be tracked where the RBAC policy must be appropriately configured to ensure that only authorized persons can access the features to revert the content to the original state.

For the F3-SF05 (Report), one (R3) of the respondents stated this feature is partly supported due to the absence of a dynamic report generator. Technically, Re-CRUD provides a report based on the current year and monthly data. However, the report is currently unable to generate customized request reports, for example reports based on a specific date range. Furthermore, three (R2, R6 and R7) respondents stated that F3-SF06 (Retention) is partly supported. Re-CRUD retention features are designed with a specific duration. Even though it has various options of retention duration (6 months, 1, 3, 5 and 7 years), these options may not match the developer requirement due to different policies by the WA stakeholder. The retention duration options are currently not flexible and require manual code modification, and may cause a repetitive task.

As mentioned earlier, most of the feature set weighting overall score is higher than the baseline score. However, one of the respondents (R3) score is 86.5% which is 1.5% lower than the baseline score since R3 is not satisfied with the F2-SF01 (RBAC), F3-SF03 (audit trail), F3-SF05 (retention) where for the F3-SF03 (audit trail) features, the SFWS is 4 (mandatory) which affect the scores. As explained in the previous paragraph, each integrated feature has a scope of functions and limitations. It may not generally be complete; however, it can still provide basic functions for each feature.

F. System Usability Evaluation Result

Eight experts have executed system usability evaluation after completing the case study for the feature analysis. This evaluation aims to capture feedback on Re-CRUD effectiveness and efficiency in web application development. Ten system usability scale (SUS) [115] questions have been asked, which focused on two aspects (effectiveness and efficiency). The index for every question was calculated from equation 3, and the results are presented in percentages. Table IX shows the collective score for effectiveness and efficiency together with the SUS score percentage.

Based on Table IX, in the effectiveness aspect (ES1, ES2, ES3, ES4, ES5), respondents strongly agree that Re-CRUD is effective in completing their task in web application development (ES1) and agreeing that it meets their requirements and is easy to use (ES4, ES5). The score for ES2 is 85%, where the developer required additional features to complete their task due to the other special requirement for a specific project. However, Re-CRUD is designed to be flexible to be applied in various types of web application projects. In the aspect of supplementing Re-CRUD with additional components (ES3), the score is 82.5%, where the respondents feel that they are still required to supplement Re-CRUD with

other components, but most of the required components are already exist in the Re-CRUD environment. The additional suggested component is automation in data/content backup, error logging, and testing from the respondent comments. Fig. 11 shows the score for Re-CRUD effectiveness. Overall, Re-CRUD effectively performs code automation for web application development.

TABLE IX. RE-CRUD EFFECTIVENESS AND EFFICIENCY SCORE

			Score					SUS Score (%)
			1	2	3	4	5	
Effectiveness (ES)	ES1	It allows me to accomplish my tasks.	0	0	0	0	8	100
	ES2	I think I would not need a system with more features for my tasks.	0	0	0	6	2	85
	ES3	I would not need to supplement Re-CRUD with an additional components.	0	0	1	5	2	82.5
	ES4	I found the system unnecessarily complex	0	0	0	1	7	97.5
	ES5	This system's capabilities would meet my requirements.	0	0	1	1	6	92.5
Efficiency (EC)	EC 1	It saves me time when I use it.	0	0	0	0	8	100
	EC 2	I found the various functions in this system were well integrated.	0	0	0	1	7	97.5
	EC 3	I tend to reduce a lot of mistakes with this system.	0	0	0	0	8	100
	EC 4	I don't make many errors with this system.	0	0	0	1	7	97.5
	EC 5	I don't have to spend a lot of time correcting things with this system	0	0	0	0	8	100

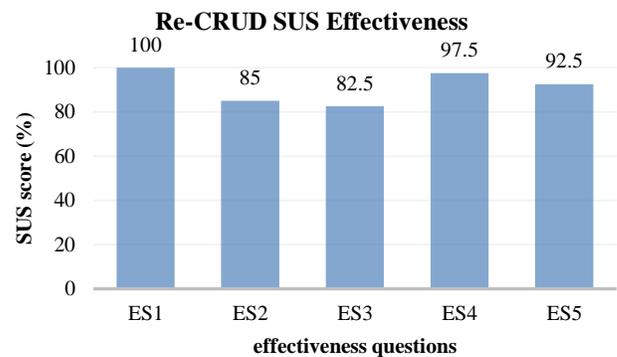


Fig. 11. Re-CRUD Effectiveness Score.

For the efficiency aspects (EC1, EC2, EC3, EC4, EC5), the respondents strongly agree with all questions in efficiency aspects. It can be seen from questions EC1 until EC5 that it has an indexed percentage above 95%. In terms of time consumption in completing respondent tasks (EC1, EC5), Re-CRUD can support rapid development, reducing the development time by providing more automation and reducing

the tendency to reduce mistakes in code writing due to human error. Re-CRUD also presented a high percentage of score in functions integration (EC2) and lower error from the default generated code automation (EC4). Fig. 12 shows the score for Re-CRUD efficiency, and it can be determined that Re-CRUD are efficient in reducing development time, systematic in features integration and reducing code error.

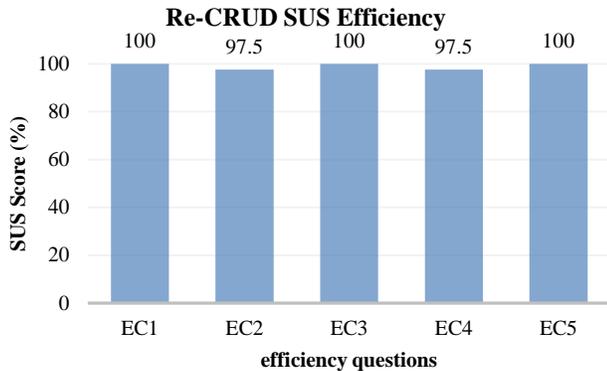


Fig. 12. Re-CRUD Efficiency Score.

V. DISCUSSION

Three types of evaluation have been conducted (case study, feature analysis and expert validation), and the results are used as the basis of these findings. The case study is used to execute EDMS development using Re-CRUD to demonstrate the code automation with the integrated important aspects. From the web application development perspective, Re-CRUD can establish the EDMS with the integrated features. The code automation from Re-CRUD can be executed appropriately in the web server environment and produces no error when performing a task. Therefore, based on the analysis of the evaluation results, Re-CRUD has reached the following conclusions without hesitation:

A. Integrated Web Application Important Features with Code Automation

The Re-CRUD output offers extra features and solutions for modern web applications. Technically the extra features can be considered must-have features since the most web application is designed to create, manage and maintain digital data or content. The records management features have been integrated into the Re-CRUD tested and compared with four other WAF CRUD generators. The feature analysis result shows that Re-CRUD has better CRUD operation output and offers more features than the other WAF CRUD generator.

B. Reduce Web Application Development Time and Code Error

With the integrated features, Re-CRUD enables the developer to speed up the development process; since all of the modern WA necessary routine features have been generated, the developer can focus on other important functions that the stakeholder requires. Using Re-CRUD also enables the developer to build the prototype and test the WA rapidly. Expert validation shows that EC1, EC3, and EC5 have the most excellent efficiency scores (100%), where Re-CRUD may

save developers time and minimize the time in code debugging due to code automation. With more code and features automation, the tendency of prone to coding error due to human mistakes also decreases.

C. Achieved Usability Attributes

Effectiveness and efficiency are the usability attributes identified for this research evaluation. An expert validation was commenced to assess the Re-CRUD with this attribute. Eight industry experts have been invited to participate in the validation process. The SUS questionnaire with 5 points Likert scale is used, and both of the usability attributes, effectiveness and efficiency result achieved the SUS acceptable score. Considering this result, there is no doubt that Re-CRUD is effectively and efficiently acceptable in handling WA development.

VI. CONCLUSION

The overall score indicates that Re-CRUD with the integrated electronic records management features is effective and efficient in developing a WA. Overall, Re-CRUD receives a perfect score for feature sets 1 and 4, while the other CRUD generators receive a 50% score for the same feature set. Re-CRUD provides integrated RBAC for feature set 2, although it is only deemed partially supported because the other CRUD generator does not provide integrated RBAC. The electronic records management capabilities emphasize Feature Set 3, and the majority of the features have been effectively integrated into Re-CRUD, with a score of 93.3%. Although other CRUD generators cover some electronic records management features, the function is only partially supported. More code modification and enhancement are required for the other CRUD outputs to incorporate key electronic records management features. Re-CRUD allows the WA developer to create a rapid WA prototype with a more practical solution to digital content management through the integration of records management tools. The integrated records management features in CRUD operation enable the developer to save more time in coding the routine code for the web application to provide the essential features such as search and reporting in the web application. It makes the development process faster with the code automation for all the routine codes. This research also has a limitation where Re-CRUD only includes eleven features with a specific task to be evaluated. Many other features can be included, but only the listed features are integrated and tested to maintain flexibility. Re-CRUD has been applied and tested on EDMS, small size of WA. A larger scale of WA may behave differently as the data table will be more prominent, and the data processing may be more complex. In future work, more features should be incorporated into Re-CRUD, i.e., progressive web apps (PWA) and improve the Re-CRUD practical ability in various web application development. At the same time, it can effectively provide mobile application ability to the generated CRUD.

ACKNOWLEDGMENT

The study was funded by the Encouragement Research Grant (Vote No. Q.K130000.3856.20J92) awarded by Universiti Teknologi Malaysia.

REFERENCES

- [1] G. J. Houben, P. Barna, F. Frascar, and R. Vdovjak, "Hera: Development of semantic web information systems," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 2722, pp. 529–538, 2003.
- [2] T. Isakowitz, M. P. Bieber, and F. Vitali, "Web Information Systems," *Commun. ACM*, vol. 41, no. 7, pp. 78–80, Jul. 1998.
- [3] A. Athanasiadis and Z. Andreopoulou, "A Web Information System Application on Forest Legislation: The Case of Greek Forest Principles," *Procedia Technol.*, vol. 8, pp. 292–299, Jan. 2013.
- [4] G. I. Papadimitriou, A. I. Vakali, G. Pallis, S. Petridou, and A. S. Pomportsis, "Simulation in Web Data Management," in *Applied System Simulation*, Springer US, 2003, pp. 179–199.
- [5] Z. Qu, S. Ninan, A. Almosa, K. G. Chang, S. Kuruvilla, and N. Nguyen, "Synoptic reporting in tumor pathology: Advantages of a web-based system," *Am. J. Clin. Pathol.*, vol. 127, no. 6, pp. 898–903, Jun. 2007.
- [6] C. Barry, "Issues and Perspectives on Web-based Information Systems Development," in *Third International Asia-Pacific Web Conference, 2000*.
- [7] M. Stauffer, *Laravel: Up and Running: A Framework for Building Modern PHP Apps*, 2nd ed. O'Reilly Media, 2019.
- [8] C. Pitt, *Pro PHP MVC*. Apress, 2012.
- [9] M. Miles, "Using web2py Python framework for creating data-driven web applications in the academic library," *Libr. Hi Tech*, vol. 34, no. 1, pp. 164–171, 2016.
- [10] Massimo Di Piero, "web2py for Scientific Applications," *Comput. Sci. Eng.*, vol. 13, no. 2, pp. 64–69, 2011.
- [11] I. Rauf and I. Porres, "Beyond CRUD," in *REST: From Research to Practice*, Springer New York, 2011, pp. 117–135.
- [12] R. Rodriguez-Echeverria, J. C. Preciado, J. Sierra, J. M. Conejero, and F. Sanchez-Figueroa, "AutoCRUD: Automatic generation of CRUD specifications in interaction flow modelling language," *Sci. Comput. Program.*, vol. 168, pp. 165–168, Dec. 2018.
- [13] S. P. Onesinus, "Laravel CRUD with bootstrap jQuery and Mysql," *OSPT*, 2019.
- [14] M. Kevin and K. McArthur, *Pro PHP: Patterns, Frameworks, Testing and More*, 1st ed. United States of America: Apress, 2008.
- [15] R. Rodriguez-Echeverria, J. M. Conejero, J. C. Preciado, and F. Sanchez-Figueroa, "AutoCRUD - Automating IFML Specification of CRUD Operations," in *Proceedings of the 12th International Conference on Web Information Systems and Technologies, 2016*, vol. 1, pp. 307–314.
- [16] J. Watts and G. Jorge, *CakePHP 2 Application Cookbook*. USA: Packt Publishing, 2014.
- [17] J. Martin, *Managing the database environment*. USA: Prentice Hall, 1983.
- [18] S. Tenzin, T. Lhamo, and Tsheten Dorji, "Design and Development of E-Commerce Web Application for Cooperative Store," *Int. Res. J. Eng. Technol.*, vol. 9, no. 2, pp. 846–846, 2022.
- [19] P. McFedries, *Web Coding & Development All-in-One For Dummies*. New Jersey: John Wiley & Sons, Inc., 2018.
- [20] Cake Software Foundation, *CakePHP Cookbook Documentation*. 2017.
- [21] Rails Guides Team, "Getting Started with Rails — Ruby on Rails Guides," 2015. [Online]. Available: https://guides.rubyonrails.org/getting_started.html#getting-up-and-running-quickly-with-scaffolding. [Accessed: 28-Nov-2019].
- [22] S. Sinha, "Introduction to Laravel," in *Beginning Laravel*, Apress, 2019, pp. 1–10.
- [23] P. Noppadon and W. Panita, "Development of a Ubiquitous Learning System with Scaffolding and Problem-Based Learning Model to Enhance Problem-Solving Skills and ICT Literacy," *Int. J. e-Education, e-Business, e-Management e-Learning*, vol. 3, no. 3, 2013.
- [24] D. Golding, *Beginning CakePHP: From novice to professional*. Apress, 2008.
- [25] R. H. Hall, A. Digennaro, J. Ward, and N. Havens, "Usability Assessment Of A Web-Based Learning System For Teaching Web Development: A Progressive Scaffolding Approach," in *Information Systems, 2003*, pp. 1–10.
- [26] R. H. Mark, *Instant CakePHP Starter*. Packt Publishing, 2013.
- [27] S. I. Ahmad, T. Rana, and A. Maqbool, "A Model-Driven Framework for the Development of MVC-Based (Web) Application," *Arab. J. Sci. Eng.*, vol. 47, no. 2, pp. 1733–1747, Feb. 2022.
- [28] R. Das and L. P. Saikia, "Comparison of Procedural PHP with Codeigniter and Laravel Framework," *Int. J. Curr. Trends Eng. Res.*, vol. 2, no. 6, pp. 42–48, 2016.
- [29] R. Dāsa, "Learn CakePHP," in *Learn CakePHP*, Berkeley: Apress, 2016.
- [30] R. Hu, Z. Wang, J. Hu, J. Xu, and X. Jun, "Agile Web development with Web framework," in *2008 International Conference on Wireless Communications, Networking and Mobile Computing, 2008*.
- [31] N. Bandirmali, "mtCMF: A novel memory table based content management framework for automatic website generation," *Comput. Stand. Interfaces*, vol. 58, pp. 43–52, May 2018.
- [32] L. Daly, *Next-generation web frameworks in Python*. O'Reilly Media, Inc, 2007.
- [33] B. Kitchenham, "DESMET: A method for evaluating Software Engineering methods and tools," 1996.
- [34] S. Murugesan, "Web Application Development: Challenges And The Role Of Web Engineering," in *Web Engineering: Modelling and Implementing Web Applications*, Springer London, 2007, pp. 7–32.
- [35] M. H. Cloyd, "Designing user-centered web applications in web time," *IEEE Softw.*, vol. 18, no. 1, pp. 62–69, Jan. 2001.
- [36] M. Y. Ivory and M. A. Hearst, "Improving web site design," *IEEE Internet Comput.*, vol. 6, no. 2, pp. 56–63, 2002.
- [37] D. A. Siegel, "The business case for user-centered design," *interactions*, vol. 10, no. 3, p. 30, May 2003.
- [38] J. Rode, M. B. Rosson, and M. Perez-Quinones, "The challenges of web engineering and requirements for better tool support," *Methods*, 2002.
- [39] S. Paul, A. Mitra, and S. Dey, "Issues and challenges in web crawling for information extraction," in *Bio-Inspired Computing for Information Retrieval Applications*, IGI Global, 2017, pp. 93–121.
- [40] D. Kunda, M. Chishimba, M. Mulenga, and V. Chama, "An Analysis of Security and Performance Concerns in Mobile Web Application Development: Challenges and Open Issues," *Int. J. Recent Contrib. from Eng. Sci. IT*, vol. 5, no. 3, pp. 26–40, Oct. 2017.
- [41] P. Mole and P. V Mole, "Progressive Web Apps: A Novel Way for Cross-Platform Development," 2018.
- [42] F. Chiti, R. Fantacci, G. Pasi, and F. Tisato, "Context-Awareness in Autonomic Communication and in Accessing Web Information: Issues and Challenges," in *Wisdom Web of Things*, Springer International Publishing, 2016, pp. 107–118.
- [43] C. Krintz, "The AppScale cloud platform: Enabling portable, scalable web application deployment," *IEEE Internet Comput.*, vol. 17, no. 2, pp. 72–75, 2013.
- [44] C. Shahabi, F. Banaei-Kashani, Y. S. Chen, and D. McLeod, "Yoda: An accurate and scalable Web-based recommendation system," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 2172, pp. 418–432, 2001.
- [45] S. Murugesan and Y. Deshpande, "Meeting the challenges of web application development," in *Proceedings of the 24th international conference on Software engineering - ICSE '02*, 2002, p. 687.
- [46] A. Ginige and S. Murugesan, "The essence of Web engineering," *IEEE Multimed.*, vol. 8, no. 2, pp. 22–25, Apr. 2001.
- [47] A. W. Anuar, N. Kama, A. Azmi, and H. M. Rusli, "Analysis And Practical Application Of Web Framework And Crud Operation For Web Application Development," in *2nd International Professional Doctorate and Postgraduate Symposium 2021*, 2021, pp. 233–237.
- [48] N. Kama, S. Basri, S. A. Ismail, and R. Ibrahim, "Using static and dynamic impact analysis for effort estimation," *IET Softw.*, vol. 10, no. 4, pp. 89–95, Aug. 2016.
- [49] C. Kai, J. Omokore, and R. K. Miller, *Practical CakePHP Projects*. USA: Apress, 2009.
- [50] H. Knublauch, R. W. Ferguson, N. F. Noy, and M. A. Musen, "The Protégé OWL Plugin: An Open Development Environment for Semantic

- Web Applications,” *Int. Semant. Web Conf.*, vol. 3298, pp. 229–243, 2004.
- [51] H. V. Nguyen, C. Kästner, and T. N. Nguyen, “Exploring variability-aware execution for testing plugin-based web applications,” in *Proceedings - International Conference on Software Engineering*, 2014, no. 1, pp. 907–918.
- [52] K. Schlegel, T. Weisgerber, F. Stegmaier, M. Granitzer, and H. Kosch, “Balloon Synopsis: A JQuery Plugin to Easily Integrate the Semantic Web in a Website?,” in *Proceedings of the 2014 International Conference on Developers - Volume 1268*, 2014, pp. 19–24.
- [53] R. Connolly, R. Hoar, S. Mukherjee, and A. K. Bhattacharjee, *Fundamentals of web development*. Pearson Education, 2015.
- [54] M. Darabseh and J. P. Martins, “Risks and Opportunities for Reforming Construction with Blockchain: Bibliometric Study,” *Civ. Eng. J.*, vol. 6, no. 6, pp. 1204–1217, Jun. 2020.
- [55] D. Ko, K. Ma, S. Park, S. Kim, D. Kim, and Y. Le Traon, “API document quality for resolving deprecated APIs,” in *Proceedings - Asia-Pacific Software Engineering Conference, APSEC*, 2014, vol. 2, pp. 27–30.
- [56] A. Austin and L. Williams, “One technique is not enough: A comparison of vulnerability discovery techniques,” in *International Symposium on Empirical Software Engineering and Measurement*, 2011, pp. 97–106.
- [57] I. Medeiros, N. F. Neves, and M. Correia, “Automatic detection and correction of Web application vulnerabilities using data mining to predict false positives,” in *WWW 2014 - Proceedings of the 23rd International Conference on World Wide Web*, 2014, pp. 63–73.
- [58] Y. Mahmood, N. Kama, and A. Azmi, “A systematic review of studies on use case points and expert-based estimation of software development effort,” *J. Softw. Evol. Process*, vol. 32, no. 7, p. e2245, Jul. 2020.
- [59] S. I. Ahamed, A. Pezewski, and A. Pezewski, “Towards framework selection criteria and suitability for an application framework,” in *International Conference on Information Technology: Coding Computing, ITCC*, 2004, vol. 1, pp. 424–428.
- [60] N. Prokofyeva and V. Boltunova, “Analysis and Practical Application of PHP Frameworks in Development of Web Information Systems,” in *Procedia Computer Science*, 2016, vol. 104, pp. 51–56.
- [61] M. Giatsoglou, V. Koutsonikola, K. Stamos, A. Vakali, and C. Zigkolis, “Dynamic code generation for cultural content management,” in *Proceedings - 14th Panhellenic Conference on Informatics, PCI 2010*, 2010, pp. 21–24.
- [62] S. A. Becker and A. Berkemeyer, “Rapid application design and testing of Web usability,” *IEEE Multimed.*, vol. 9, no. 4, pp. 38–46, Oct. 2002.
- [63] Shari Thurow and Nick Musica, “Understanding Search Usability,” in *When Search Meets Web Usability, USA: New Riders*, 2009, pp. 2–14.
- [64] K. C. Haggerty and R. E. Scott, “Do, or Do Not, Make Them Think?: A Usability Study of an Academic Library Search Box,” *J. Web Librariansh.*, vol. 13, no. 4, pp. 296–310, Oct. 2019.
- [65] M. Matera, F. Rizzo, and G. T. Carughi, “Web usability: Principles and evaluation methods,” in *Web Engineering*, Springer Berlin Heidelberg, 2006, pp. 143–180.
- [66] T. Brinck and E. Hofer, “Automatically evaluating the usability of web sites,” in *CHI '02 extended abstracts on Human factors in computing systems - CHI '02*, 2002, p. 906.
- [67] Z. T. Shasha and M. Weideen, “Usability Measurement of Web-based Hotel Reservation Systems,” in *1st TESA International Conference*, 2016, no. September, pp. 1–14.
- [68] V. Thoma and J. Dodd, “Web Usability and Eyetracking,” in *Eye Movement Research*, Springer, Cham, 2019, pp. 883–927.
- [69] T. Felin and S. Kauffman, “The Search Function and Evolutionary Novelty,” *SSRN Electron. J.*, Oct. 2019.
- [70] B. Porebski, K. Przystalski, L. Nowak, and L. N. Bartosz Porebski, Karol Przystalski, *Building PHP Applications with Symphony, CakePHP, and Zend Framework*, 1st ed. GBR: Wiley Publishing, 2011.
- [71] T. Gipp and J. Ebert, “Functional web applications,” in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2007, vol. 4607 LNCS, pp. 194–209.
- [72] A. Gutlić and E. Mujčić, “Intelligent Web Application for Search of Restaurants and Their Services,” in *Networks and Systems*, vol. 83, Springer, 2020, pp. 452–469.
- [73] M. A. Jiménez-Crespo, “What is (not) web localization in translation studies,” *J. Int. Localization*, vol. 3, no. 1, pp. 38–60, Aug. 2016.
- [74] B. Esselink, “The Evolution of Localization,” in *Guide to Localization*, 2003, pp. 21–29.
- [75] Jiménez-Crespo and Miguel A., *Translation and web localization*. London: Routledge, 2013.
- [76] P. Zhu, “Language Problems to Be Coped with in Web Localization,” *J. Tech. Writ. Commun.*, vol. 39, no. 1, pp. 57–78, Jan. 2009.
- [77] J. Williams, “Correctly assessing the ‘ilities’ requires more than marketing hype,” *IT Prof.*, vol. 2, no. 6, pp. 65–67, Nov. 2000.
- [78] J. Hvorecký and M. Beňo, “Linkages between PISA Results and E-Working,” *Emerg. Sci. J.*, vol. 5, no. 3, pp. 294–304, Jun. 2021.
- [79] B. Tambaip, I. H. Wayangkau, S. Suwarjono, M. Loupatty, A. F. Adam, and H. Hariyanto, “Preservation of the Muyu Indigenous Language with an Android-based Dictionary,” *Emerg. Sci. J.*, vol. 4, no. 0, pp. 85–101, Oct. 2021.
- [80] S. Murugesan and A. Ginige, “Introduction and Perspectives,” in *Web Engineering, India: Idea Group Inc*, 2005.
- [81] A. W. Anuar, N. Kama, A. Azmi, and H. M. Rusli, “A Multivocal Literature Review on Records Management Potential Components in CRUD Operation for Web Application Development,” *Int. J. Model. Simulation, Sci. Comput.*, Apr. 2022.
- [82] Ira A. Penn and Gail B. Pennix, “Records inventory,” in *Records Management Handbook*, 2nd ed., New York: Routledge, 2017.
- [83] Patricia C. Franks, “Records Retention Strategies Inventory Appraisal Retention and Disposition,” in *Records and Information Management*, American Library Association, 2013, pp. 84–114.
- [84] Judith Read and Mary Lea Ginn, “Electronic Records Management,” in *Records Management*, 10th ed., Boston: Cengage Learning, 2016.
- [85] UK National Archive, “Records Management retention scheduling,” United Kingdom, 2012.
- [86] UK National Archive, “Born-digital records and metadata,” *Information Management-Digital Record Transfer*, 2017. [Online]. Available: <https://www.nationalarchives.gov.uk/information-management/management-information/digital-records-transfer/what-are-born-digital-records/>. [Accessed: 06-Dec-2020].
- [87] M. Diamond, “How to Implement a Record Retention Schedule for Electronic and Other Records,” *Association of Corporate Counsel (ACC)*, 2017. [Online]. Available: <https://www.acc.com/resource-library/how-implement-record-retention-schedule-electronic-and-other-records#>. [Accessed: 20-Feb-2021].
- [88] C. A. Lee, “Computer-Assisted Appraisal and Selection of Archival Materials,” in *Proceedings - 2018 IEEE International Conference on Big Data, Big Data 2018*, 2019, pp. 2721–2724.
- [89] R. Harvey and D. Thompson, “Automating the appraisal of digital materials,” *Libr. Hi Tech*, vol. 28, no. 2, pp. 313–322, 2010.
- [90] A. Vellino and I. Alberts, “Assisting the appraisal of e-mail records with automatic classification,” *Rec. Manag. J.*, vol. 26, no. 3, pp. 293–313, 2016.
- [91] IRMT, “Managing the Creation, Use and Disposal of Electronic Records,” *Int. Rec. Manag. Trust*, 2009.
- [92] M. Crockett, “User Guide to Retention and Disposal Schedules Council of Europe Records Management Project,” 2011.
- [93] N. W. Lo, C. Y. Wu, and Y. H. Chuang, “An authentication and authorization mechanism for long-term electronic health records management,” in *Procedia Computer Science*, 2017, vol. 111, pp. 145–153.
- [94] T. M. Masenya, “Application of modern technologies in the management of records in public libraries,” *J. South African Soc. Arch.*, vol. 53, pp. 65–79, Dec. 2020.
- [95] H. Guo, W. Li, M. Nejad, and C. C. Shen, “Access control for electronic health records with hybrid blockchain-edge architecture,” in *Proceedings - 2019 2nd IEEE International Conference on Blockchain, Blockchain 2019*, 2019, pp. 44–51.

- [96] B. Oladejo and S. Hadzidedić, "Electronic records management – a state of the art review," *Rec. Manag. J.*, vol. ahead-of-p, no. ahead-of-print, Feb. 2021.
- [97] P. Joseph, S. Debowski, and P. Goldschmidt, "Search behaviour in electronic document and records management systems: An exploratory investigation and model," *Inf. Res.*, vol. 18, no. 1, 2013.
- [98] V. L. Lemieux, "Trusting records: is Blockchain technology the answer?," *Rec. Manag. J.*, vol. 26, no. 2, pp. 110–139, Jul. 2016.
- [99] J. Namukasa, "Records management and procurement performance: A case of NAADS program in the central region of Uganda," *Rec. Manag. J.*, vol. 27, no. 3, pp. 256–274, 2017.
- [100] I. Pappel, S. Butt, I. Pappel, and D. Draheim, "On the specific role of electronic document and record management systems in enterprise integration," in *Advances in Intelligent Systems and Computing*, vol. 1184, Springer Science and Business Media Deutschland GmbH, 2021, pp. 37–51.
- [101] M. Broussard and K. Boss, "Saving Data Journalism: New strategies for archiving interactive, born-digital news," *Digit. Journal.*, vol. 6, no. 9, pp. 1206–1221, Oct. 2018.
- [102] International Council on Archives, "Principles and functional requirements for records in electronic office environments," 2013.
- [103] The National Archives United Kingdom, "Migrating information between records management systems," pp. 1–35, 2017.
- [104] L. Chen, W. K. Lee, C. C. Chang, K. K. R. Choo, and N. Zhang, "Blockchain based searchable encryption for electronic health record sharing," *Futur. Gener. Comput. Syst.*, vol. 95, pp. 420–429, Jun. 2019.
- [105] J. Lengstorf and K. Wald, *Pro PHP and jQuery*. Apress, 2016.
- [106] J. Duarte, C. F. Portela, A. Abelha, J. Machado, and M. F. Santos, "Electronic Health Record in Dermatology Service," in *ENTERprise Information Systems*, 2011, pp. 156–164.
- [107] H. Hedberg and J. Lappalainen, "A preliminary evaluation of software inspection tools, with the DESMET method," in *Proceedings - International Conference on Quality Software*, 2005, vol. 2005, pp. 45–52.
- [108] C. Marshall, P. Brereton, and B. Kitchenham, "Tools to support systematic reviews in software engineering: A feature analysis," in *ACM International Conference Proceeding Series*, 2014, pp. 1–10.
- [109] F. Faiqunisa, E. Nugroho, and P. I. Santosa, "A Model of Electronic Document Management System for Limited Partnership," *J. Telemat. Informatics*, vol. 1, no. 2, pp. 69–79, Jun. 2013.
- [110] F. D. Butterfoss, V. Francisco, and E. M. Capwell, "Choosing Effective Evaluation Methods," *Health Promot. Pract.*, vol. 1, no. 4, pp. 307–313, 2000.
- [111] E. Umamaheswari and D. K. Ghosh, "Software quality: Dual experts opinion and conditional based aggregation method," *Int. J. Eng. Technol.*, vol. 6, no. 2, pp. 1167–1175, 2014.
- [112] I. Blotenberg and A. Richter, "Validation of the QJIM: A measure of qualitative job insecurity," *Work Stress*, vol. 34, no. 4, pp. 406–417, Oct. 2020.
- [113] M. Al-aaidroos, N. Jailani, and M. Mukhtar, "Expert validation on a reference model for e-auctions that conform to Islamic trading principles," *J. King Saud Univ. - Comput. Inf. Sci.*, vol. 31, no. 1, pp. 62–71, Jan. 2019.
- [114] PMI, *A Guide to the Project Management Body of Knowledge (PMBOK Guide)*, 4th ed., vol. 40, no. 2. Project Management Institute, Inc., 2008.
- [115] J. Brooke, "SUS: A quick and dirty usability scale," United Kingdom, 1986.
- [116] J. R. Lewis, "The System Usability Scale: Past, Present, and Future," *Int. J. Hum. Comput. Interact.*, vol. 34, no. 7, pp. 577–590, Jul. 2018.
- [117] S. Ratnawati, L. Widianingsih, N. Angraini, I. Marzuki Shofi, N. Hakiem, and F. Eka M Agustin, "Evaluation of Digital Library's Usability Using the System Usability Scale Method of (A Case Study)," in *2020 8th International Conference on Cyber and IT Service Management, CITSM 2020*, 2020.
- [118] K. Finstad, "The usability metric for user experience," *Interact. Comput.*, vol. 22, no. 5, pp. 323–327, Sep. 2010.
- [119] A. Bangor, P. Kortum, and J. Miller, "Determining what individual SUS scores mean: adding an adjective rating scale," *J. usability Stud.*, vol. 4, no. 3, pp. 114–123, 2009.
- [120] S. C. Peres, T. Pham, and R. Phillips, "Validation of the system usability scale (sus) in the wild," in *Proceedings of the Human Factors and Ergonomics Society*, 2013, pp. 192–196.
- [121] I. M. Pageh, A. A. J. Permana, and K. Suranata, "Usability testing and the social analysis on online counselling system for recommendations in technical vocational schools," in *Journal of Physics: Conference Series*, 2021, vol. 1810, no. 1, p. 12022.
- [122] K. Finstad, "The system usability scale and non-native English speakers," *J. Usability Stud.*, vol. 1, no. 4, pp. 185–188, 2006.
- [123] M. A. Diefenbach, N. D. Weinstein, and J. O'reilly, "Scales for assessing perceptions of health hazard susceptibility," *Health Educ. Res.*, vol. 8, no. 2, pp. 181–192, Jun. 1993.
- [124] E. P. Cox, "The Optimal Number of Response Alternatives for a Scale: A Review," *J. Mark. Res.*, vol. 17, no. 4, p. 407, Nov. 1980.
- [125] J. Nielsen, "Usability 101: Introduction to Usability," *Research-Based User Experience*, 2012. [Online]. Available: <https://www.nngroup.com/articles/usability-101-introduction-to-usability/>. [Accessed: 01-Sep-2020].
- [126] A. Holzinger, "Usability engineering methods for software developers," *Communications of the ACM*, vol. 48, no. 1, pp. 71–74, 01-Jan-2005.
- [127] J. Nielsen, "Usability metrics: Tracking interface improvements," *IEEE Softw.*, vol. 13, no. 6, pp. 12–13, 1996.
- [128] L. Faulkner, "Beyond the five-user assumption: Benefits of increased sample sizes in usability testing," in *Behavior Research Methods, Instruments, and Computers*, 2003, vol. 35, no. 3, pp. 379–383.
- [129] C. W. Turner, J. R. Lewis, and J. Nielsen, "Determining Usability Test Sample Size," 2006.

A Novel Readability Complexity Score for Gujarati Idiomatic Text

Jatin C. Modh¹

Research Scholar
Gujarat Technological University
Ahmedabad, India

Jatinderkumar R. Saini^{2*}

Symbiosis Institute of Computer
Studies and Research, Symbiosis
International (Deemed University)
Pune, India

Ketan Kotecha³

Symbiosis Centre for Applied Artificial
Intelligence, Symbiosis International
(Deemed University)
Pune, India

Abstract—Gujarati language is used for conversation by more than 55 million people worldwide and it is more than 1000 years old language. It is the chief language of the Indian state of Gujarat. There are many dialects of Gujarati like Standard Gujarati, Amdawadi Gujarati, Kathiawadi Gujarati, Kutchi Gujarati etc. The Gujarati language is very rich in morphology like other Indo-Aryan languages like Hindi. Many readability tests are available in the English language, but no readability complexity test is available for the Gujarati idiomatic text. The Complexity score is the sub concept of the readability test. In order to define complexity level of Gujarati text, complexity score of Gujarati text is calculated. We deployed a novel readability complexity score calculation method in which we considered the number of letters of each word, the number of diacritics of each word, Gujarati idiomatic text of n-gram where n=1 to 9, Gujarati idiomatic text of m-meaning idioms where m=1 to 7. The complexity score is calculated as the sum of word complexity score, diacritics complexity score, n-gram complexity score of Gujarati idioms and m-meaning complexity score of Gujarati idioms. We emphasized Gujarati idiomatic text for the calculation of complexity score as idioms make the text more complex to understand. This is an innovative and first of its kind work in the research community of Gujarati language. The results are hopeful enough to employ the suggested complexity score method for developing a readability test method for natural language processing tasks for the Gujarati language.

Keywords—Complexity; Gujarati; idiomatic text; natural language processing (NLP); readability

I. INTRODUCTION

Gujarati language is named after the people of Gurjar people who are said to have established in the middle of the 5th century CE. Gujarati language is used by more than 55 million people worldwide and it is more than 1000 years old language based on Indo-Aryan languages. Gujarati language stands in 26th position among the most spoken native language in the world. Gujaratis are spread all over the world. It is the chief language of the Indian state of Gujarat. It is also main language in the union territories of Daman and Diu, Dadra and Nagar Haveli. Outside of India, it is spoken all over the world in many countries like United States, Canada, UK, Southeast African countries etc. There are many dialects of Gujarati like Standard Gujarati, Amdawadi Gujarati, Kathiawadi Gujarati, Kutchi Gujarati etc. The spelling of Gujarati words is based on pronunciation [1][2].

A. Gujarati Script

Gujarati is written similar to the Devanagari script except it does not have the horizontal line above characters. The Gujarati alphabet has mainly 34 consonants, 13 vowels and 10 digits working as a building block of the Gujarati language. Sarth Gujarati dictionary consists more than 65000 words excluding technical or slang words [3]. Gujarat vowels and Gujarati consonants can be written as independent letters or by combining with diacritic marks. Diacritics play a very important role in building meaningful words and thus vocabulary of the Gujarati language. Fig. 1 shows the use of diacritics with the letter **દ**. Gujarati diacritics and conjuncts make Gujarati script more effective for written and communication purposes [4][5].

B. Gujarati idioms

An idiom is a group of words but whose meaning is established by the usage and not as the literal meaning of its separate words. Gujarati people are using Gujarati idioms for expressing thoughts, feelings and messages. Gujarati idioms are not understandable for non-Gujarati people as well as for children of a lower standard. Gujarati idioms can be understood by the surrounding context information [6]. Gujarati idioms can be classified on the base of N-grams and on the base of the number of m-meanings [8]. Gujarati idioms can also be classified as static idioms versus inflected idioms. Here we consider idioms as unfamiliar words. Example of Gujarati idiom is જાલ લેવું 'jala levum' i.e. to take a vow. It is bigram/2-gram and single-meaning idiom.

C. Text Complexity

English language consists of 26 alphabets with 21 consonants and 5 vowels for writing. Generally, three aspects are used to decide the complexity of the English text: quantitative measures, qualitative measures and concerns involving to the reader and task [7]. The Gujarati language is morphologically very rich compared to the English language. The Gujarati language consists of 18 diacritics [6]. Diacritics make many possible word formations by suffixing or prefixing any letter. Using diacritics various inflectional forms are possible for Gujarati verbs and Gujarati nouns [9]. Here only quantitative measures are considered for complexity as our text is just in written form. Factors such as sentence, word length and the frequency of unfamiliar words are used as quantitative measures of text complexity.

*Corresponding Author

Independent vowels	અ	આ	ઇ	ઈ	ઉ	ઊ	એ	ઐ	ઓ	ઔ	અં	અઃ	ઋ
	a	aa	i	ee	u	oo	e	ai	o	au	am	Ah	ru
Common Diacritics →		◌ા	◌િ	◌ી	◌ુ	◌ૂ	◌ે	◌ૈ	◌ો	◌ૌ	◌ં	◌ઃ	◌ૃ
દ + Diacritics →		દા	દિ	દી	દુ	દૂ	દે	દૈ	દો	દૌ	દં	દઃ	દૃ

Fig. 1. Use of Diacritics in the Building Gujarati Conjuncts with Letter d.

The rest of the paper is organized as follows: Section II corresponds to the literature review related to text complexity and Gujarati text; Section III represents the methodology including collection of idioms data and the method of calculating Gujarati text complexity; Section IV covers the results and analysis; finally, the limitations, conclusion and future work are represented in Section V.

II. RELATED LITERATURE REVIEW

A readability score is computer calculated score which roughly decides what level of knowledge needed by someone to be able to read a text easily. Various researches have been performed for the study of the readability and complexity of the various languages. Various work related to readability formula have been carried out.

Harvey [7] represented three-part model for measuring text complexity namely qualitative measures, quantitative measures and reader & task. Quantitative measures consider more lexile level text as more complex than less lexile text. A qualitative factor considers layout, text structure, language features, purpose and meaning etc descriptors. Reader & task is dependent on the professional judgment of teachers about the complex text. Author used a Rubric - a set of guidelines to decide the complexity of the English text.

Uccelli [10] considered parameters like word length, frequency of unfamiliar terms, sentence length and text cohesion for the quantitative dimension of the complexity of English language text. The author emphasized that multiple themes, multiple perspectives, content-specific knowledge, figurative or ambiguous language make English text very complex text.

Anet [11] defined text complexity as easy or hard text in terms of reading based on qualitative and quantitative text features. Important quantitative parameters for defining text complexity are structure, meaning or purpose, language and knowledge requirement for particular English text.

Barge [12] calculated the English text complexity Rubric using 10 dimensions; each dimension can receive a score between 0 and 10 to indicate the optimal benefit for students. 100 points is the best possible overall score for a text and interpreted collective text scores depend on the different points. The rubric provides a framework to assist educators.

Flesch and Kincaid [13] designed readability tests to indicate the difficulty of English passages to understand. They represented two tests namely Flesch Reading-Ease and Flesch-Kincaid Grade level. Same core measures of sentence length and word length are used by the authors for the two tests.

Tillman and Hagberg [14] used Swedish and English language to test the compatibility of readability algorithms.

They tested three algorithms namely Coleman-Liau index (CLI), Lasbarhetsindex (LIX) and Automated Readability Index (ARI) on Wikipedia articles. Authors concluded that CLI seem to perform less well on higher level text but works excellent on the Bible like easy to read text in Swedish and English languages, whereas LIX and ARI work on average as well as hard texts in both Swedish and English languages.

Venugopal et al. [15][16] analyzed the complex words in Hindi language sentences and experimented with whether classical readability parameters of the English language can be applied to the Hindi language or not for determining the complexity of the word. They demonstrated that the frequency parameter plays an important role in determining the complexity of a word in Hindi sentence. As per their study, the length of a word is not a significant factor; the number of syllables plays an important predictor of word complexity. Researchers used five tree-based ensemble models out of a total of eight classifiers to extract the important features.

Sinha et al. [17] presented that the English readability formulas are not helpful for Hindi and Bangla languages. They proposed two new readability models for Hindi text documents and Bangla text documents. They customized standard structural parameters like word length, sentence length, number of syllables/word, number of polysyllabic words, number of consonant-conjuncts and number of polysyllabic words per 30 sentences.

Mehta and Majumder [18] explored large-scale media text of three Indo-Aryan languages Gujarati, Bengali, and Hindi as a part of quantitative analysis. As per their statistical study of the corpus, Bengali piece of writing might be more difficult to read than Hindi or Gujarati; Gujarati corpus has more diversity in vocabulary and it contains double type-token ratio than that of Bengali; Hindi is less artificial compare to Gujarati but more compared to Bengali, etc.

Modh and Saini [19][20] collected 2-gram to 9-gram Gujarati idioms and classified them as single-meaning to seven-meaning idioms based on a number of meanings. Authors [6] detected Gujarati idioms from the entered text using diacritics and suffix-based rules. Researchers [8] also exploited IndoWordNet for deciding the meaning of idioms on the base of surrounding contextual information.

Based on this exhaustive literature assessment and evaluation, English language text is analyzed by many researchers in detail for deciding the readability score of the English text by applying different standard parameters. Indo-Aryan languages like Hindi, Bengali and Gujarati are analyzed by some researchers by comparing it with English parameters. Very less work is done specially for Gujarati language text. No researchers have calculated the readability complexity score of

the Gujarati idiomatic text and No other researchers have tried to identify Gujarati idioms from the Gujarati text.

The paper highlights on the study of the complexity of Gujarati text by considering parameters like the number of letters in the individual word and the number of diacritics of the individual word. This paper also considers the presence of idioms in the text and also considers the type of idioms in the text and decides the complexity level of the Gujarati text. The extent of this paper is to analyze letters, diacritics, words and idioms within Gujarati text. This deployment helps in the study of the complexity of Gujarati idiomatic text.

III. METHODOLOGY

For the calculation of the complexity score of Gujarati text, four parameters are considered (1) the number of letters of each word (2) the number of diacritics of each word (3) the number of Gujarati idioms. If Gujarati idioms are found in the text, then the idiom(s) are classified in two ways: N-gram classification and M-meaning classification. Different complexity points are allocated to different classifications of idioms. The complexity score is calculated as the summation of meaning complexity, gram complexity, word complexity and diacritics complexity.

Complexity Score=Meaning Complexity Score + Gram Complexity Score + Word Complexity Score + Diacritics Complexity Score

A. Collection of Data

By and large 3472 distinct Gujarati idioms are accumulated from different Gujarati language resources [21][22]. Idiom data collection is basically for the recognition of Gujarati idioms from the Gujarati text.

B. N-Gram Idiom Classification and Complexity Points

Idioms are classified on the basis of N-gram model. Idioms can be classified as 2-gram or bigram, trigram or 3-gram, 4-gram or four-gram, 5-gram, 6-gram, 7-gram, 8-gram, 9-gram.

Idiom up to 9-gram was found. 1-gram idioms are specific personage idioms that represent the historical or fictional special character identity in a play. Example of 7-gram Gujarati idiom is રાન રાન ને પાન પાન થઈ જવું 'rana rana ne pana pana thai javum' i.e. getting into a bad situation.

Table I shows the classification of idioms on the base of N-grams and their corresponding complexity point calculation method. Bigrams and trigrams are more in number, so both are getting relatively more complexity points compared to other N-gram idioms.

C. M-Meaning Idiom Classification and Complexity Points

Idioms are also classified on the base of their meanings. Gujarati Idiom has a single meaning or more than one meaning. For single meaning idioms, a dictionary based approach is used to understand the meaning of an idiom, but for multiple meaning idioms, surrounding contextual information is needed to understand the idiomatic text. So it is complex to understand multiple-meaning idioms. So M-meaning idioms, corresponding M-complexity points are assigned. Table II shows the classification of M-meaning idioms and corresponding complexity points for the calculation of the complexity score. Gujarati Idioms are found from single meaning to seven meaning idioms. More complexity points are assigned for 7-meaning idioms as it requires more effort to understand by studying the surrounding contextual text.

For example ઠેકાણું કરવું 'thek anum karavum' is a 7-meaning idiom as it has 7 different possible meanings depending upon the context like ઉપયોગમાં લેવું 'upayogamam levum' i.e. to use, કન્યાને સારે ઘેર પરજાવવી 'kanyane sare ghera paranavavi' i.e. marry the bride to the right person, કાસળ કાઢવું 'kasala kadhavum' i.e. to kill, ખલાસ કરવું 'khalasa karavum' i.e. use-up, છેવટની ક્રિયા કરવી 'chevatani kriya karavi' i.e. take the last action, મારીને દાટી દેવું 'marine dati devum' i.e. kill and bury, યોગ્ય સ્થાને ગોઠવી દેવું 'yogya sthane gothavi devum' i.e. arrange in the right place.

TABLE I. COMPLEXITY POINT CALCULATION FOR EACH N-GRAM IDIOM

Sr. No.	N-gram Idioms	Count	(Count/Total Idioms) *10	Complexity Point (Roundup to 2 decimal)
1	Unigrams	58	0.167050691	0.17
2	Bigrams	2102	6.054147465	6.06
3	Trigrams	992	2.857142857	2.86
4	4-Grams	244	0.702764977	0.71
5	5-Grams	63	0.181451613	0.19
6	6-Grams	9	0.025921659	0.03
7	7-grams	2	0.005760369	0.01
8	8-grams	1	0.002880184	0.01
9	9-grams	1	0.002880184	0.01
	Total Idioms	3472		

TABLE II. COMPLEXITY POINT TABLE FOR M-MEANING IDIOMS

Sr. No.	M-meaning idioms	Count	Number of meaning(s)	Complexity Point
1	single-meaning	1806	1	1
2	2-meanings	953	2	2
3	3-meanings	504	3	3
4	4-meanings	193	4	4
5	5-meanings	13	5	5
6	6-meanings	1	6	6
7	7-meanings	2	7	7
	Total Idioms	3472		

D. Diacritics Complexity Score

If there are no diacritics in the Gujarati word, then the particular word is considered simple and easy to read. For example, Gujarati word રમઝમ 'ramzam' i.e. ramzam has no diacritics. Another example of a Gujarati word, ચાદર 'chadar' i.e. sheet has 1 diacritics. If there are more diacritics in the particular word, then the particular word is difficult to read. If the count of diacritics of a particular word is 0 or 1, then that particular word is considered as simple, so 0 complexity point is assigned. If the count of diacritics of a particular word is 2, then 0.2 complexity point is assigned. If the count of diacritics of a particular word is 3 or 4, then 0.5 complexity point is assigned. If the count of diacritics of a particular word is 5 or 6, then 1 complexity point is assigned. If the count of diacritics of a particular word is greater than or equal to 7, then 2 complexity point is assigned. Table III shows the complexity point table on the base of number of diacritics of a particular word.

E. Word Complexity Score

If the count of letters of a particular word is 1, 2 or 3, then that word is considered as simple, so 0 complexity point is assigned. If the count of letters of a particular word is 4 or 5, then 0.5 complexity point is assigned. If the count of letters of a particular word is 6 or 7, then 1 complexity point is assigned. If the count of letters of a particular word is greater than or equal to 8, then a 2 complexity point is assigned. Table IV shows the complexity point table on the base of the number of letters of a particular word.

F. Database of Idioms

An Idiom database is required to store the collected Gujarati idioms. This idiom database is used to identify idioms from the input text to decide the complexity of the Gujarati idiomatic text. Idiom column stores the base form of the idiom in the idiom database. Fields like idiom, Gujarati meaning of idiom, English meaning of idiom and other related fields are created as a part of the Idiom database [6][23].

TABLE III. COMPLEXITY POINT TABLE ON THE BASE OF NUMBER OF DIACRITICS OF PARTICULAR WORD

Sr. No.	No. of diacritics of particular word	Complexity Point	Example
1	0	0	રમઝમ 'ramzam' i.e. ramzam
2	1	0	ચાદર 'chadar' i.e. sheet
3	2	0.2	વાદળી 'vadali' i.e. blue
4	3 to 4	0.5	ચાદરમાં 'chadarman' i.e. in the sheet
5	5 to 6	1	ચીડિયાપણું 'chidiyapanum' i.e. irritability
6	Greater than or equal to 7	2	પ્રતિદ્વંદ્વિત્વ 'pratidhvandhita' i.e. competition

TABLE IV. COMPLEXITY POINT TABLE ON THE BASE OF NUMBER OF LETTERS OF PARTICULAR WORD

Sr. No.	Number of letters of particular word	Complexity Point	Example
1	1 to 3	0	આકાશ 'aakash' i.e. sky
2	4 to 5	0.5	બતાવવી 'batavavi' i.e. showing
3	6 to 7	1	પ્રયોજનભૂત 'prayojanbhut' i.e. purposeful
4	Greater than or equal to 8	2	તત્વજ્ઞાનીઓનો 'tatvagnaniono' i.e. of philosophers

G. Proposed Model

Fig. 2 explains the steps for the proposed algorithm/model.

Step 1: Accept the Gujarati text from the user.
Step 2: Pre-processing step 2.1: Eliminate whitespaces from starting and ending side of the text 2.2: Eliminate all whitespaces in between the text
Step 3: Tokenize all the words of entered text.
Step 4: Eliminate Gujarati stop words from the entered text.
Step 5: Find out Gujarati idioms from the entered text using the idiom database
Step 6: Calculate the gram-complexity score for idioms as per Table I.
Step 7: Calculate the meaning-complexity score for idioms as per Table II.
Step 8: Count the number of letters of individual word
Step 9: Count the number of diacritics of individual word
Step 10: Calculate diacritics complexity score as per Table III.
Step 11: Calculate word complexity score as per Table IV.
Step 12: Calculate complexity score=Gram-complexity score + Meaning-complexity score + Diacritics complexity score + Word complexity score
Step 13: Display complexity level results of Input text.

Fig. 2. Algorithm for the Proposed Model.

The entered input is the Gujarati text which may or may not contain any unfamiliar words, including the Gujarati idioms. The output will be the analysis of Gujarati text with complexity score, which takes into consideration various factors, and the corresponding complexity level.

IV. RESULT AND ANALYSIS

Gujarati text containing zero or more idioms is given as an input and output shows the related complexity score and complexity level of the inputted Gujarati text. The algorithm ignores the stop words in calculating complexity scores. Output also shows the stop words found in the input text. It also displays total words, total stop words, total letters, and total diacritics used in the input Gujarati text. It calculates Gram complexity score, meaning complexity score, diacritics complexity score and word complexity score as per weight defined in Table I, Table II, Table III and Table IV. The proposed model implements Table V for showing the complexity type or complexity level as an output.

We now present a few examples for the execution of the proposed algorithm for calculating the novel complexity score for the different instances of the Gujarati text. In Example 1, Example 2 and Example 3, different Gujarati text is given as an input. In Example 1, the input text is taken from the standard 1 Gujarati textbook. The output confirms that the complexity type of the text is SIMPLE. This is expected for the text used for teaching the first graders in the age group of generally 5 to 6 years.

TABLE V. COMPLEXITY SCORE INTERPRETATION TABLE

Sr. No.	Complexity Score	Complexity Type	Notes
1	0.0-20.0	SIMPLE	Very easy words.
2	20.0-40.0	FAIRLY SIMPLE	Fairly Easy.
3	40.0-60.0	MEDIUM	Medium complexity
4	60.0-80.0	COMPLEX	Complex
5	80.0 or more	VERY COMPLEX	Extremely complex

Example1:

<p>INPUT TEXT=વરસાદ આવે રમઝમ વાદળી ચાલે ઝમઝમ, મોટા મોટા છાંટા પડતા આભથી એ નીચે સરતા. આકાશ આજે ધમધમ વીજ ચમકતી ચમચમ, ધરની લીલી ચાદર ઓઢે લીલી ચાદરમાં ધરની પોઢે. ઢમઢમ ઢોલ વગડાવો, વરસાદને સૌ વધાવો. ‘varasada ave ramajhama vadali cale jhamajhama, mota mota chanta padata abhathi e nice sarata. akasa gaje dhamadhama vija camakati camacama, dharati lili cadara odhe lili cadaramam dharati podhe. dhamadhama dhola vagadavo, varasadane sau vadhavo.’</p> <p>OUTPUT: STOP WORDS FOUND----> આવે, એ, નીચે, ‘ave, e, nice,’</p> <p>Total Words in input Text: 34 Total Idioms Found: 0 Meaning Complexity Score: 0 Gram Complexity Score: 0 Word Complexity Score: 6 Diacritics Complexity Score: 3.2 Total letters in input: 97 Total diacritics in input: 41 Total stop words in input: 3 Complexity Score = 9.2 Complexity Type = SIMPLE</p>
--

In Example 2, the input text contains the collection of 13 idioms. Output identifies these 13 idioms and from these 13 idioms, 8 idioms are with 1-meaning, 3 idioms are with 2 meanings, 1 idiom with 3 meanings and 1 idiom with 4 meanings. Output also identifies different N-gram wise idioms. Corresponding meaning complexity score and gram complexity score are calculated. Word complexity score and Diacritics complexity score is also calculated. Finally, the complexity score is calculated and the complexity type is decided on the base of the range of complexity score.

Example2:

<p>INPUT TEXT=એક કાને સાંભળી બીજે કાને કાઢી નાખવું ઢ સંસાર માંડવો આગ લાગવી અક્કલ ચરવા જવી આંખમાં પાણી આવવું આકાશ પાતાળ જેટલું અંતર આંખ બતાવવી અક્કડ ને અક્કડ રહેવું જમીન પર પગ ન મૂકવો નાક ઉપર માખી ન બેસવા દેવી એકે પથ્થર ઉથામ્યા વગરનો ન રહેવો રાત કહે તો રાત દહાડો કહે તો દહાડો ‘eka kane sambhali bije kane kadhi nakhavum dha sansara mandavo aga lagavi akkala carava javi ankhamam pani avavum akasa patala jetalum antara ankha batavavi akkada ne akkada rahevum jamina para paga na mukavo naka upara makhi na besava devi eke paththara uthamya vagarano na rahevo rata kahe to rata dahado kahe to dahado’</p> <p>OUTPUT: STOP WORDS FOUND----> એક, જેટલું, ને, રહેવું, પર, ન, ઉપર, ન, ન, તો, તો, ‘eka, jetalum, ne, rahevum, para, na, upara, na, na, to, to,’</p> <p>Total Words in input Text: 53 Total Idioms Found: 13</p> <p>8 Idioms With 1 Meaning(s) 3 Idioms With 2 Meaning(s) 1 Idioms With 3 Meaning(s) 1 Idioms With 4 Meaning(s) Meaning Complexity Score: 21</p> <p>1 Idioms With 8 Gram(s) 1 Idioms With 7 Gram(s) 2 Idioms With 6 Gram(s) 1 Idioms With 5 Gram(s) 2 Idioms With 4 Gram(s) 2 Idioms With 3 Gram(s) 3 Idioms With 2 Gram(s)</p>

1 Idioms With 1 Gram(s)
Gram Complexity Score: 26.5

Word Complexity Score: 3.5
Diacritics Complexity Score: 5.9

Total letters in input: 114
Total diacritics in input: 66
Total stop words in input: 11

Complexity Score = 56.9
Complexity Type = **MEDIUM**

In Example 3, the complexity score is calculated as 75.3, which is in the range of 60.0-80.0, so the output of the complexity type is COMPLEX.

Example3:

INPUT TEXT=અંતરવેદના અંધામૂંઘી અતિસૌરભ હાથ ઝાલ અનુસંધાન અવસન્નતા અવસન્નત્વ આશોકિત આદીનવ આમ્રવૃક્ષ ઈંદ્રશસ્ત્ર ઈંદ્રાયુધ ઈંમ્નિહાન ઉપદ્રવ ત્રિદશાંકુશ ત્રિદશાયુધ ધાતુરાજક નિરીક્ષણ પરિચારક પરેશાની પર્યેષણ પિકવલ્લભ પૂછપરછ પ્રતિકુળ પ્રિયાંબુ ભોગવિલાસ અંતર રાખ મજ્જાસ્સ મનોવ્યથા મુશ્કેલી મેઘજ્યોતિ મેઘભૂતિ રતિકલલ રતિકેવિ રતિસંહતિ રતિસુખ વજ્રાશનિ વસંતદૂત વસંતદ્રુ વસંતદ્રુમ વિટંબાણ વિરુદ્ધતા વિષયભોગ વિષયસુખ વ્યાકુલપાણું વ્યાકુળતા શતકોટી સહાયરૂપ સૌદામની સૌદામિની સ્ત્રીગમન સ્ત્રીસંસારી સ્ત્રીસુખ સ્ત્રીસેવન હેરાનગત

'antaravedana andhadhundhi atisaurabha hatha jhala anusandhana avasannata avasannatva ajnankita adinava amravrksa indrasastra indrayudha imtihana upadrava tridasankusa tridasayudha dhaturajaka niriksana paricaraka paresani paryesana pikavallabha puchaparacha pratikula priyambu bhogavilasa antara rakha majjarasa manovyatha muskeli meghajyoti meghabhuti ratikalaha ratikeli ratisanhati ratisukha vajrasani vasantaduta vasantadru vasantadruma vitambana virudhdhata visayabhoga visayasukha vyakulapanum vyakulata satakoti sahayarupa saudamani saudamini strigamana strisansarga strisukha strisevana heranagata'

OUTPUT:
STOP WORDS FOUND---->
Total Words in input Text: 57
Total Idioms Found: 2

1 Idioms With 2 Meaning(s)
1 Idioms With 5 Meaning(s)
Meaning Complexity Score: 7

2 Idioms With 2 Gram(s)
Gram Complexity Score: 12.2

Word Complexity Score: 33
Diacritics Complexity Score: 23.1

Total letters in input: 278
Total diacritics in input: 163
Total stop words in input: 0

Complexity Score = 75.3
Complexity Type = **COMPLEX**

V. CONCLUSION, LIMITATIONS AND FUTURE WORK

The proposed Gujarati text complexity prediction model was successfully implemented and it was based on the number of diacritics of the individual word, the number of letters of the individual word and on the number of idioms. Different complexity points are considered on the basis of N-gram idioms and M-meaning idioms. Gujarati idioms are considered as unfamiliar words to understand the Gujarati text. The complexity score of Gujarati text is calculated as the

summation of diacritics complexity points, word complexity points, N-gram idiom complexity points and M-meaning idiom complexity points.

The proposed model could not recognize idioms those are not stored in the idiom database for assigning complexity points. Future work is to assemble all Gujarati idioms to correct this drawback. In the future enhancement of the model, particular domain vocabulary can be used for defining complexity levels.

Based on the outcome achieved, it is advocated that the projected readability complexity score calculation method is worth implementing in the real world for the community of Gujarati language. To the best of our knowledge, it is the first and novel readability complexity score calculation method and complexity type prediction method for the Gujarati Idiomatic text. The proposed method considers the Gujarati idioms as unfamiliar words and assigns weightage accordingly by dynamically detecting them from the input text. The proposed method opens the path for other Gujarati language researchers in defining readability levels for Gujarati text as well as natural language processing tasks for the Gujarati language.

REFERENCES

- [1] Wikipedia, "Gujarati language", https://en.wikipedia.org/wiki/Gujarati_language (accessed March 23, 2022).
- [2] Yourdictionary, "Gujarati Language Overview and Common Words"; Available Online: <https://reference.yourdictionary.com/other-languages/gujarati-language-words.html> (accessed March 23, 2022).
- [3] GujaratiLexicon, "Let's Learn Gujarati"; Available Online: <http://www.letslearngujarati.com/vowels> (accessed March 23, 2022).
- [4] Audichya M. and Saini J.R., 2019, "An Overview of Optical Character Recognition for Gujarati Typed and Handwritten Characters", Available online: <https://www.researchgate.net/publication/350173104>.
- [5] Rakholia R.M. and Saini J.R., 2015, "The Design and Implementation of Diacritic Extraction Technique for Gujarati Written Script using Unicode Transformation Format", proc. of IEEE International Conference on Electrical, Computer and Communication Technologies (ICECCT-2015), Coimbatore, India, vol. 2, pages 654-659, Available online: <https://ieeexplore.ieee.org/document/7226037>.
- [6] Modh J.C. and Saini J.R., "Dynamic Phrase Generation for Detection of Idioms of Gujarati Language using Diacritics and Suffix-based Rules", International Journal of Advanced Computer Science and Applications (IJACSA), 12(7), 2021. <http://dx.doi.org/10.14569/IJACSA.2021.0120728>.
- [7] Harvey S., "A Beginner's Guide to Text Complexity", Generation Ready, 2013; Available online: <https://www.generationready.com/wp-content/uploads/2021/04/Beginners-Guide-to-Text-Complexity.pdf>.
- [8] Modh J.C. and Saini J.R., "Using IndoWordNet for Contextually Improved Machine Translation of Gujarati Idioms", International Journal of Advanced Computer Science and Applications (IJACSA), 12(1), 2021. <http://dx.doi.org/10.14569/IJACSA.2021.0120128>.
- [9] Kapadia U. and Desai A., "Rule Based Gujarati Morphological Analyzer", IJCSI International Journal of Computer Science Issues, Volume 14, Issue 2, March 2017, <https://www.ijcsi.org/papers/IJCSI-14-2-30-35.pdf>.
- [10] Uccelli P., "Why do so many adolescents struggle with content-area reading?", Available Online: <https://iris.peabody.vanderbilt.edu/module/sec-rdng2/cresource/q1/p02/> (accessed March 23, 2022).
- [11] The Achievement Network Ltd, "Text Complexity", Available Online: <https://www.achievementnetwork.org/anetblog/eduspeak/text-complexity> (accessed March 23, 2022).
- [12] Barge J., "Common Core Georgia Performance Standards Text Complexity Rubric", Georgia Department of Education, 2011; Available

- online: <https://www.gpb.org/sites/default/files/2020-06/handout-1-ccgps-ela-textcomplexity-guide.pdf>.
- [13] Wikipedia, "Flesch–Kincaid readability tests", Available Online: https://en.wikipedia.org/wiki/Flesch–Kincaid_readability_tests (accessed March 23, 2022).
- [14] Tillman R. and Hagberg L. (2014), "Readability algorithms compability on multiple languages", Digitala Vetenskapliga Arkivet (DiVA), Stockholm, Available Online: <https://www.diva-portal.org/smash/get/diva2:721646/FULLTEXT01.pdf>.
- [15] Venugopal G., Dhanya P., Saini J.R. (2021), "Analyzing Complex Words in Hindi using Parameters of Classical Readability Formulae (Part 1)", Computer Science Journal of Moldova, 29(3):366-387. Online: [http://www.math.md/files/csjm/v29-n3/v29-n3-\(pp366-387\).pdf](http://www.math.md/files/csjm/v29-n3/v29-n3-(pp366-387).pdf).
- [16] Venugopal G., Dhanya P., Saini J.R. (2022), "Revisiting the Role of Classical Readability Formulae Parameters in Complex Word Identification (Part 2)", Computer Science Journal of Moldova, 30(1):49-63. Available Online: [http://www.math.md/files/csjm/v30-n1/v30-n1-\(pp49-63\).pdf](http://www.math.md/files/csjm/v30-n1/v30-n1-(pp49-63).pdf).
- [17] Sinha M., Sharma S., Dasgupta T. and Basu A. (2012), "New Readability Measures for Bangla and Hindi Texts", Proceedings of COLING 2012: Posters, pages 1141–1150. Available Online: <https://aclanthology.org/C12-2111.pdf>.
- [18] Mehta P. and Majumder P. (2014), "Large Scale Quantitative Analysis of three Indo-Aryan Languages", Journal of Quantitative Linguistics, Available Online: https://www.researchgate.net/publication/272496714_Large_Scale_Quantitative_Analysis_of_three_Indo-Aryan_Languages.
- [19] Modh J.C. and Saini J.R., 2018, "A Study of Machine Translation Approaches for Gujarati Language", International Journal of Advanced Research in Computer Science, Volume 9, No. 1, January-February 2018, pages 285-288; Available online: ijarcs.info/index.php/Ijarcs/article/download/5266/4497.
- [20] Modh J.C. and Saini J.R., "Context Based MTS for Translating Gujarati Trigram and Bigram Idioms to English," 2020 International Conference for Emerging Technology (INCET), 2020, pp. 1-6, doi: 10.1109/INCET49848.2020.9154112.
- [21] GujaratiLexicon, Gujaratilexicon.com, Available online: <http://www.letslearngujarati.com/about-us> (accessed March 23, 2022).
- [22] Rudhiprayog ane kahevatsangrah, published by Director of Languages, Gujarat State, Gandhinagar. 2010.
- [23] Saini J.R. and Modh J.C., "GIdTra: A dictionary-based MTS for translating Gujarati bigram idioms to English," 2016 Fourth International Conference on Parallel, Distributed and Grid Computing (PDGC), 2016, pp. 192-196, doi: 10.1109/PDGC.2016.7913143.

Importance of Memory Management Layer in Big Data Architecture

Maha Dessokey¹, Sherif M. Saif², Hesham Eldeeb³
Computer and Systems Department
Electronics Research Institute
Cairo, Egypt

Sameh Salem⁴, Elsayed Saad⁵
Computer and Systems Department
Faculty of Engineering, Helwan University
Cairo, Egypt

Abstract—The generation of daily massive amounts of heterogeneous data from a variety of sources presents a challenge in terms of storage and analysis capabilities and brings new problems into high-performance computing clusters. To better utilize this huge and heterogeneous data, the continuous development of advanced Big Data platforms and Big Data analytic techniques are required. One of the significant issues with in-memory Big Data processing platforms, such as Apache Spark, is the user’s responsibility to decide whether the intermediate data should be cached or not. In addition, the data may be kept in several storage systems and physically scattered over different racks, regions, and clouds. Data need to be close to the computation nodes and hence data locality issue is a challenge. In this paper, using a distinct memory management layer between the data processing layer and the data storage layer, which automatically caches data without the need for any interaction from the applications’ developers, is evaluated. K-means, PageRank and WordCount workloads from the HiBench benchmark beside a real case to predict the price of Real Estate that is implemented using Gradient Boosting Regression Tree model, are used to evaluate this framework. Experiments show that the memory management layer outperforms the Apache Spark in reducing the execution time.

Keywords—Apache Spark; Big Data; data analytics algorithms; memory management

I. INTRODUCTION

For both academic, business and engineering communities, Big Data analytics for storing, processing, and analysing large scale heterogeneous datasets has become a must-have tool. New Big Data analysis techniques [1], as well as the constant development of advanced Big Data platforms [2], are required to take full advantage of this massive and heterogeneous data. Fig. 1 shows the Big Data architecture. The architecture consists of four layers.

Data storage layer, which contains multiple storage systems such as distributed file systems Hadoop Distributed File System, GlusterFS, Ceph, etc, or remote access file systems such as Amazon S3, Swift, Google Cloud Storage, etc. and it can contain tools and techniques such as relational databases and NoSQL tools.

Resource management layer controls resource management, scheduling, and security. Examples of resource managers are YARN, Mesos, and Kubernetes K8s.

Data processing layer, which contains one of the Big Data platforms such as the open-source Hadoop Map Reduce [3], the Apache Spark platform [4], The Apache Flink, etc.

Application layer, the top layer, can contain any application type, such as batch, graph analytics, machine learning, streaming, etc.

The default Big Data architecture has many challenges [5][6], such as:

- In some Big Data platforms that support in-memory computing, such as Apache Spark and Apache Flink, developers can cache data that will most likely be reused. As a result, it is entirely up to the developer to decide which data to cache in memory. Determining which data should be cached is a difficult task when dealing with jobs that consist of operations with complex dependencies. When there isn’t enough memory, caching all data in memory will result in a significant performance loss. While disc caching saves RAM, it reduces the efficiency of in-memory computing;
- Data locality is another problem for data processing; data are physically scattered over different racks, regions, and clouds. Data must be near to where data computation occurs to be processed;
- The data needed by the application may be stored in different storage systems, the application developer must be aware of all Storage APIs, such as HDFS API, FUSE API, S3 API, REST API.

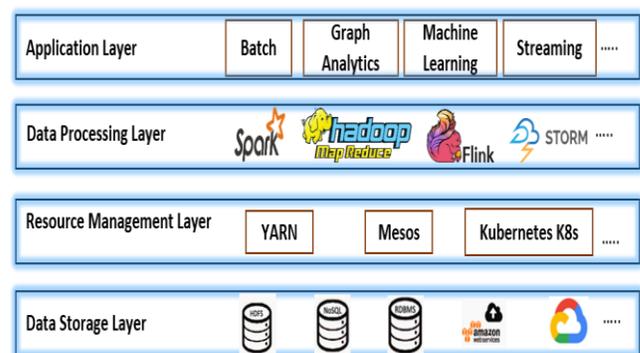


Fig. 1. The Big Data Architecture.

All the previous challenges necessitate the use of a distinct memory management layer between the data processing layer and the data storage layer.

In this paper, to evaluate the distinct memory management layer, the Apache Spark platform is used in the processing layer because Apache Spark boosts Hadoop's performance by up to 100x using in-memory cluster computing [7] and it is widely used in a variety of application domains, including bioinformatics [8], image processing [9], deep learning [10], finance [11], and astronomy [12], etc. The next section gives a background about Apache Spark and memory management layer. Three workloads from HiBench [13] are used for evaluation with randomly generated dataset, then a real case study to predict the price of real estate with real data set is developed to be used in the evaluation.

The rest of the paper is organized as follows. In Section 2 an overview about Apache Spark and memory management layer is given. In Section 3, a review of the related work is presented. In Section 4, the experimental setup and workloads are de-scribed. In Section 5, our experimental results are discussed. Finally, the conclusion of our findings is in Section 6.

II. BACKGROUND

In this section, an overview is given on Apache Spark which is used in data processing layer, Apache Spark Standalone cluster manager which is used in resource management layer, Hadoop Distributed File System which is used in data storage layer and the memory management layer.

A. Apache Spark

Apache Spark [14] is a computing engine and a suite of libraries for processing data in parallel on computer clusters. Apache Spark is one of the most used open-source engines for Big Data processing. Apache Spark is compatible with several popular programming languages (Python, R, Scala, and Java). It offers libraries for a wide range of operations, including data loading and SQL queries, as well as machine learning and streaming computation. The basic abstraction in Apache Spark is a Resilient Distributed Dataset (RDD). It acts as an immutable, partitioned collection of elements. These elements can be run in parallel.

The Apache Spark cluster can be managed by Apache Spark Standalone cluster manager or by other cluster managers as Mesos or YARN. The Apache Spark Standalone cluster manager is used to test Apache Spark performance in our experimental setting. As shown in Fig. 2, the Apache Spark master receives the application then a driver process is created and connected to the cluster manager through the SparkContext [4]. The cluster manager allocates the Apache Spark workers. Each worker contains an executor processes. The driver process is in charge of running the main () function, keeping track of the Apache Spark application's progress, responding to a user's program input, and analyzing, distributing, and scheduling work throughout the executors. The executors are in charge of completing the job that has been given to them and reporting the state of the computation back to the driver node.

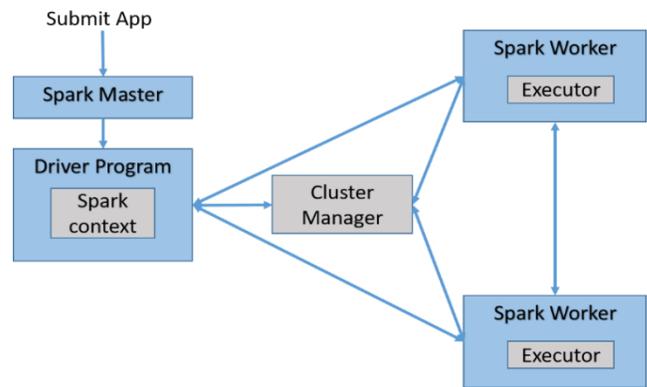


Fig. 2. Apache Spark Cluster Architecture.

Apache Spark was created to read and write data from and to Hadoop Distributed File System (HDFS) [15], allowing it to be used with Hadoop clusters. HDFS as a distributed file system allows users to access application data quickly. It allows enormous amounts of structured and unstructured data to be managed. HDFS is a file system that divides the processing of massive data sets across inexpensive hardware clusters.

HDFS has a primary NameNode, which keeps track of where the file is kept in the cluster and multiple of DataNodes on a commodity hardware cluster. Splitting huge files into little sections known as blocks is one of the major HDFS characteristics. These blocks hold a specific amount of data that can be read and written. The block size is set to 128 MB by default. Hadoop splits up blocks and distributes them across different nodes called DataNodes.

Another main characteristic in HDFS is replication which duplicates data blocks to provide fault tolerance and allows an application to select the number of replicas for a file. The replication factor can be specified when the file is created and can be changed later. If a node fails, you can still access the data on other nodes in the HDFS cluster that have a copy of the same data. By default, HDFS duplicates blocks three times.

B. Memory Management Layer

The challenges in the main Big Data architecture, as demonstrated in the introduction, necessitate the use of a distinct memory management layer between the data processing layer and the data storage layer. Fig. 3 shows the distinct memory management layer's location in the Big Data architecture.

This distinct memory management layer:

- Caches the most frequently used data automatically. A local storage space is set aside in each node of the processing cluster for hot and transient data. This storage could be of any type (memory, SSD or HDD). The size and type of storage are determined by the user. When an application attempts to read data that is only available in shared storage, the data are duplicated in local storage. When the local storage is full, one of the eviction policies [5] can be used to determine which data should be deleted.

- Handles the distributed storage system; because the data are duplicated in local storage, this can address the data locality issue in distributed storage systems.
- Serves as a global namespace, a global namespace is an important feature of a distributed file system that makes it easy to find and access data from multiple storage systems using a single control and administration layer. The global namespace can be considered as a global file directory that allows all data from several storage systems to seem as if they were stored in a single storage system. It automatically converts the standard client-side interface to any storage interface.

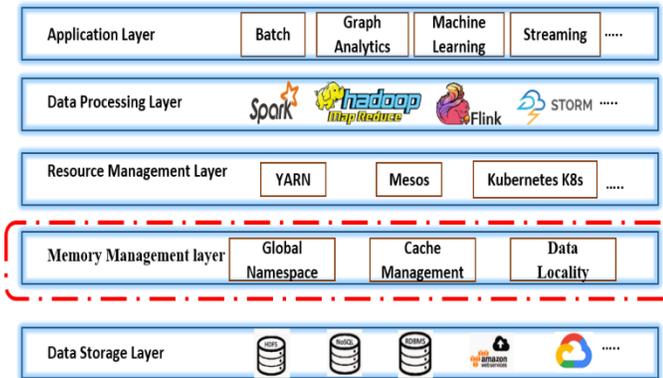


Fig. 3. The Big Data Architecture with a Distinct Memory Management Layer.

III. RELATED WORK

Many optimization strategies have been presented to improve memory management in Big Data frameworks. For Apache Spark, when the memory used to cache data reaches its capacity limit, data must be selected to be deleted to make way for new ones. Apache Spark’s cache replacement strategy uses Least Recently Used (LRU) criteria to determine which RDDs should be replaced, different cache replacement techniques were investigated [5] to improve Apache Spark performance. Apache Ignite [16], a high-performance, distributed in-memory computing platform for large-scale data sets has a cache management feature that keeps data in RAM as much as possible, having minimal interaction with the disk, but researchers in [17] observed that Apache Spark outperform Apache Ignite as Apache Ignite does not distribute well data among available nodes and it does not balance well the communication between the nodes. Other researchers are interested in investigating Apache Spark’s performance on various disk types. Doppio [18] proposed an I/O Aware performance study for Apache Spark, which measured the I/O impact of using hard disk drives (HDDs) and Solid-state drives (SSDs) with different combinations, and discovered the relationship between computation and I/O access by changing the CPU core number. As a result, the model could be used to locate the best configuration on the public cloud. Instead of reserving running memory for caching, RubiX [19], an open source project employs SSDs. It is utilized in the Azure HDInsight data caching service, which increases the performance of Apache Spark processes [20]. Delta Lake [21] by Databricks is another open source storage layer. Which also

leverages nodes’ local storage for caching, and the data is cached automatically anytime a file has to be retrieved from a remote site, resulting in much faster reading speeds. The user is not required to take any action throughout the caching process. Open Cache Acceleration Software [22], works with node memory to build a multilevel cache that optimizes system memory usage and automatically chooses the appropriate cache level for active data, allowing programmes to run quicker than they would on SSDs alone.

IV. EXPERIMENTAL SETUP

In this section, the used Big Data architecture is shown in Fig. 4, and fully described it in the following subsection then the used workloads and the implemented application are described.

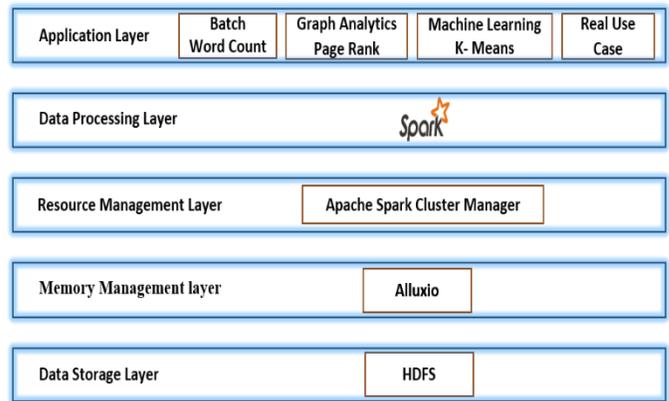


Fig. 4. Implemented Cluster’s Architecture.

A. Cluster’s Architecture

The experiments were deployed in a cluster at Electronics Research Institute. The Apache Spark cluster contains five servers, which is configured as one master and four slaves. There are 160 CPU cores and 640 GB of RAM in the cluster. The cluster services are shown in Fig. 5, and its specifications are listed in Table I. The data are stored on HDFS with data nodes on the same slaves.

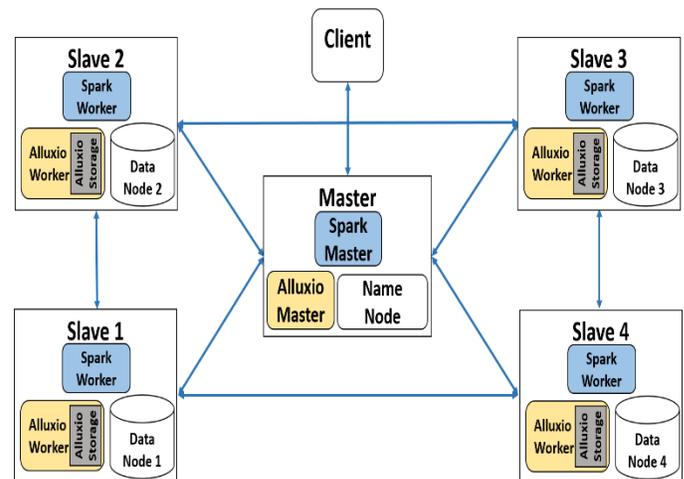


Fig. 5. Cluster’s Services.

TABLE I. CLUSTER'S SPECIFICATIONS

Servers Configuration	Processor: Intel(R) Xeon(R) CPU E5-2680 0 @ 2.70GHz Main memory:128MB Local storage:1 TB CPU cores: 32
Software	Operating system: Red Hat 4.8.3-9 JDK: 1.8 Hadoop:2.7.1 Spark: 2.4.6 Alluxio: 2.3.0
Workload	HiBench 7.1.1

In the experimental setup, Alluxio [23] a virtual distributed storage platform, is used in the memory management layer. The master and workers of Alluxio are running on the same Apache Spark cluster.

Alluxio enables users to integrate their data across multiple platforms. As shown in Fig. 6, Alluxio is made up of three different components: masters, workers, and clients. All user requests and file system metadata modifications are served by the Alluxio Master. The Alluxio Job Master is a lightweight scheduler for file system activities that are then executed on Alluxio Job Workers.

A specific amount of local Alluxio storage is determined for each Alluxio worker to store hot and transient data. Client requests to read or write data are fulfilled by Alluxio workers by reading or constructing new blocks within their local resources.

There are three scenarios for reading data:

- If the requested data are already in the worker local storage, then the client will read the file directly via the local file system (Local Cache Hit).
- If requested data are stored in another worker, the client will perform a remote read from that worker that does have the data. After the client finishes reading the data, it creates a copy locally for future reads (Remote Cache Hit).
- If the data are not available in any of the running workers, the client will read the data from the storage (Cache Miss).

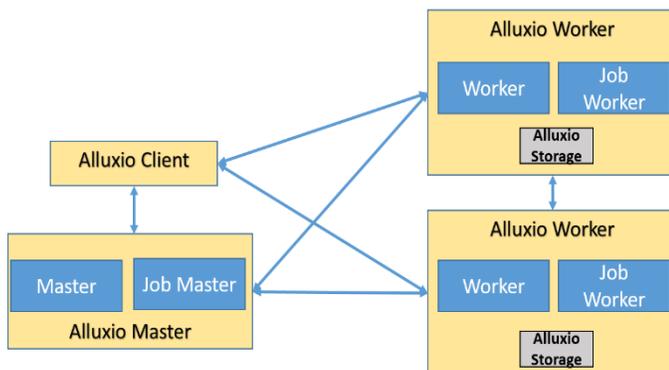


Fig. 6. Alluxio Architecture.

Workers are only in charge of managing blocks; the master is the only one who keeps track of the file-to-block mapping. Because RAM has a finite capacity, blocks in a worker may be evicted if space is full. Eviction policies as least recently used and least frequently used can be used by workers to pick which data to keep in the Alluxio space.

In our experiments some parameters must be set in order to make the Apache Spark applications access Alluxio, the Alluxio client jar must be in the classpath of all Apache Spark drivers and executors so spark.driver.extraClassPath and spark.executor.extraClassPath parameters were added to spark/conf/spark-defaults.conf and was set to the path where the alluxio-2.3.0-client.jar is located. In order for HiBench to access Alluxio, hibench.hdfs.master parameter had been set to alluxio://{Alluxio_master_Hostname}:19998 in Hibench/conf/hadoop.conf, and the Hibench/sparkbench/assembly/target/sparkbench-assembly.jar had been copied to /spark/jars.

The following configuration had been added to hadoop/coresite.xml

```
<configuration>
  <property>
    <name>fs.alluxio.impl</name>
    <value>alluxio.hadoop.FileSystem</value>
  </property>
</configuration>
```

The experiments compared between Apache Spark framework with HDFS and Apache Spark framework with a distinct memory management layer and with different RAM size per worker.

B. Workloads

For a comprehensive evaluation, first three applications from HiBench, including K-Means as a machine learning algorithm for clustering workload, Page Rank as a graph analytics workload and WordCount as a batch processing workload were used with randomly generated dataset. Those workloads were chosen to compare Apache Spark cluster performance with and without memory management layer because of their distinct properties. Then a real use case was implemented to predict real estate sale price using Gradient-Boosted Trees as a machine learning algorithm for regression.

1) *K-Means*: K-Means [24] is a popular unsupervised machine learning clustering algorithm for data mining and knowledge discovery. The idea of the algorithm as shown in Fig. 7 that it divides a collection of samples into K groups or clusters, In Fig. 7, K is equal to 3. In the initialization, the algorithm defines K centroids randomly, and makes iterations of calculations to define new centroids in order to minimize the Euclidean distances between the points forming each cluster and its centroid. The iterations are repeated until the most optimum centroids are defined or the maximum number of iterations is reached. The input data in our experiment were 100,000,000 samples generated by GenKMeans dataset based on uniform distribution and Gaussian distribution. Based on

each sample's attributes, the algorithm assigns each sample to one of the k groups iteratively. The algorithm's input parameters used in the experiments were 5 clusters, 20 dimensions, and 5 iterations.

2) *PageRank*: PageRank [25] is an iterative graph analytics algorithm, as shown in Fig. 8, it ranks items based on the number and quality of their links. The PageRank's mathematics is completely general and may be used to any graph or network in any domain. As a result, PageRank and its variations are widely used in social and information network analysis, link prediction, and recommendation systems. It's even used in road network systems analysis [26]. The PageRank algorithm used in our experiments is implemented in apache Spark MLlib. The input data were generated from web data whose hyperlinks follow the Zipfian distribution. In the experiments, the input data parameters used were 5,000,000 pages and 3 iterations.

3) *WordCount*: The WordCount workload is a batch processing workload. It scans the input data once and counts how many times each word appears.

Random text writer generates the input data, the input data size used in the experiments was 30 GB.

4) *Real estate sale price prediction*: The prediction of real estate sale price in the presence of a large number of variables is a known problem, there are many models that were built with different methods to estimated house price by inputting house features [27] [28]. The data set used in this experiment has been downloaded from [29] and was collected from some popular portals for the sale of real estate in the period from 2018 to 2021. The dataset contains 5,477,005 records with 11 features. Fig. 9 presents the output of Python code to describe the dataset fields.

Where building type parameter could be { "1" for Panel, "2" for Monolithic, "3" for Brick, "4" for Blocky, "5" for Wooden and 0 for other type}. Object type indicates the apartment type and it could be { "1" for old buildings and "2" for new building}. Level indicates apartment floor. Levels indicates the number of stores in the building. Rooms indicates the number of living rooms and if the value is "-1" then it is for studio apartment. In our implementation, the data were read from HDFS once and once from Alluxio system. The code was written in Python. The dataset first passed through cleaning phase to remove any null values or negative values in the price column and take the log value of the price. Then the correlation matrix shown in Fig. 10 has been calculated to pick the most correlated features with the price.

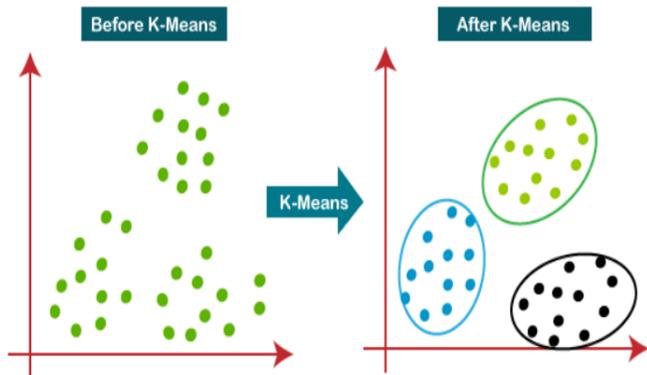


Fig. 7. K-Means Algorithm.

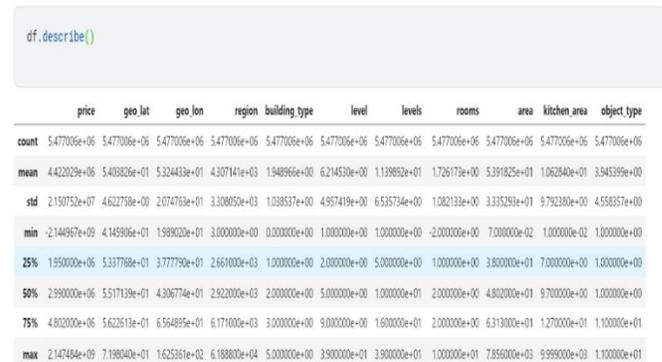


Fig. 9. Real Estate Data Set Parameters.

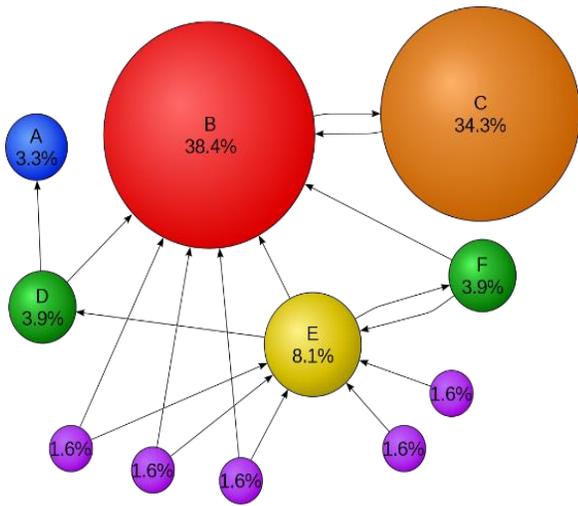


Fig. 8. Page Rank Algorithm.

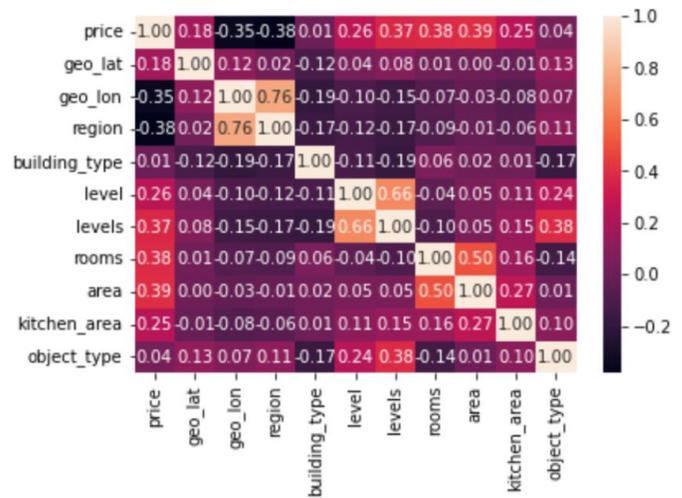


Fig. 10. Correlation Matrix.

The data set was split randomly into training set and testing set with the ratio 70:30 respectively. Gradient-Boosted Trees (GBTs) learning algorithm [30] was used to predict the price. GBTs is one of the most powerful and frequently used algorithms by data scientists for building predictive models [31]. It acts as a machine learning technique for regression and classification problems. In our application, the algorithm was used as a regression technique using pyspark.ml.regression library. Given the input variables, regression analysis estimates the conditional expectation of the price variable. The results shown in the next section were for 5 iterations, 10 max depth. To evaluate the prediction model performance, the Root Mean Squared Error (RMSE) evaluation measure was used which considers the sample standard deviation difference between the predicted and real values. The lower the RMSE, the more accurate the model predictions will be. The average RMSE in all the runs was 0.13.

V. RESULT AND DISCUSSION

In this section, the results obtained after running the experiments are shown and evaluated. The execution time in minutes is used to calculate performance measures. In the experiment, the difference between utilizing the Apache Spark with HDFS and Apache Spark using a distinct memory management layer with Alluxio was evaluated.

The results illustrated how Apache Spark with Alluxio in the memory management layer improves the performance in all cases. Fig. 11 shows the execution time of running Real Estate application, K-means, PageRank, and WordCount algorithms on ERI cluster. The worker's RAM size varies between 4, 8, 16, 32 and 64 GB to study the effect of the memory management layer with respect to the RAM size.

It can be seen from the results that the memory management layer has better performance with smaller RAM size.

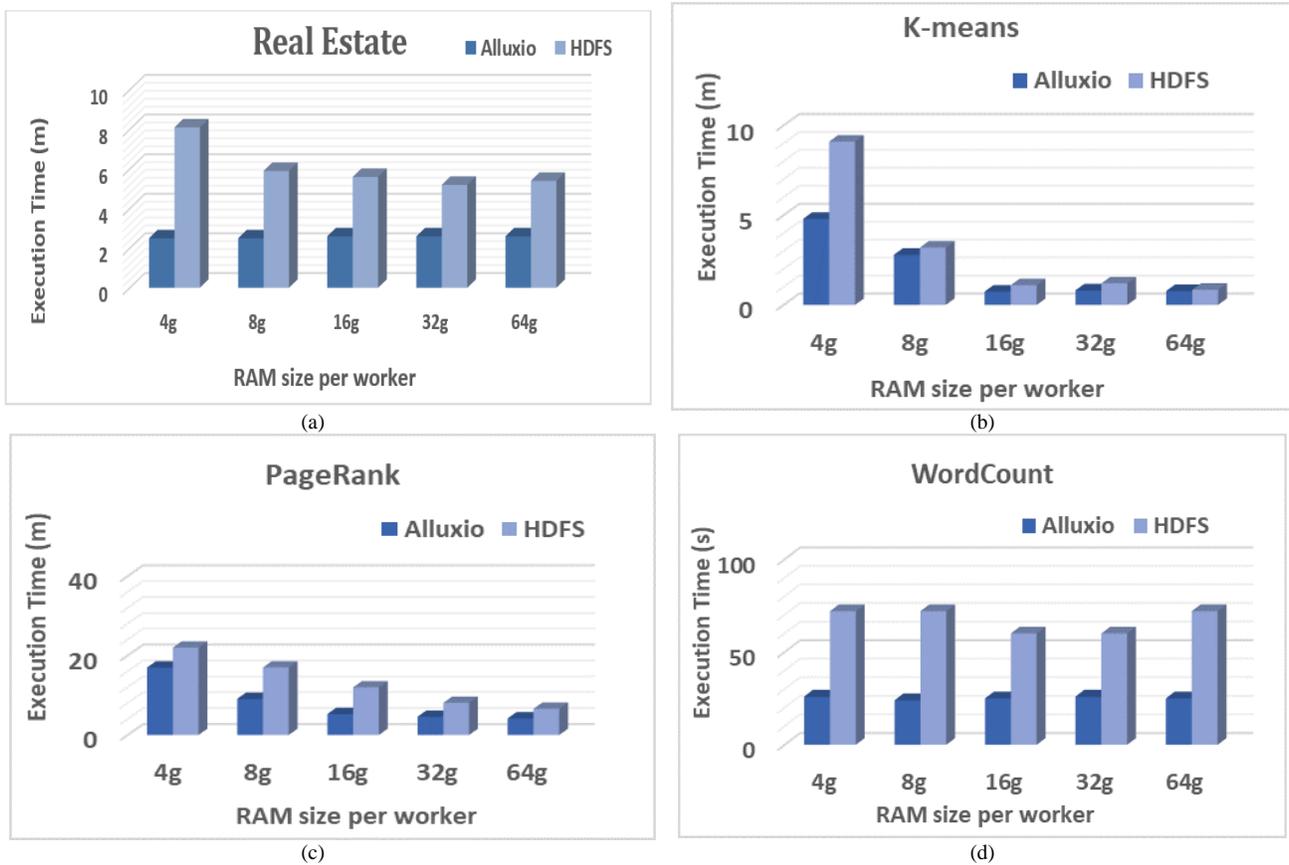


Fig. 11. The Execution Time of Running (a) Real Estate, (b) K-means; (c) PageRank and (d) WordCount on Apache Spark with HDFS and Apache Spark using Alluxio.

VI. CONCLUSION

In this paper, the impact of using a distinct memory management layer between the data processing layer and the data storage layer in Big Data Architecture was evaluated. This layer automatically caches the most frequently used data and handles the distributed storage system without the need for any interaction from the applications' developers.

First a HiBench benchmark was used with k-means, PageRank and WordCount algorithms with randomly generated workloads. Then a real case study with real dataset was implemented to predict real estate price using Gradient Boosting Regression tree. The results showed that when using distinct memory management layer, the execution time of the real estate application is up to three times faster than the normal case. So using memory management layer helps the applications' developers to get better performance up to three times faster with less effort. As a future work, other evaluations can be done with other tools, mentioned in section 3, other than Alluxio, such as Apache Ignite and RubiX.

ACKNOWLEDGMENTS

The authors send their acknowledgement to the Electronics Research Institute (ERI), Cairo, Egypt for running the experiments on ERI system.

REFERENCES

- [1] M. S. Mahmud, J. Z. Huang, S. Salloum, T. Z. Emara, and K. Sadatdiynov, "A survey of data partitioning and sampling methods to support big data analysis," *Big Data Mining and Analytics*, vol. 3, no. 2, Art. no. 2, 2020.
- [2] A. H. Ali, "A survey on vertical and horizontal scaling platforms for big data analytics," *International Journal of Integrated Engineering*, vol. 11, no. 6, Art. no. 6, 2019.
- [3] "Apache Hadoop." <https://hadoop.apache.org/>.
- [4] "Apache Spark." <https://spark.apache.org/>.
- [5] M. Dessokey, S. M. Saif, S. Salem, E. Saad, and H. Eldeeb, "Memory Management Approaches in Apache Spark: A Review," *Proceedings of the International Conference on Advanced Intelligent Systems and Informatics 2020*, Cham, 2021, pp. 394–403.
- [6] S. Lee, J. -Y. Jo and Y. Kim, "Survey of Data Locality in Apache Hadoop," *IEEE International Conference on Big Data, Cloud Computing, Data Science & Engineering (BCD)*, 2019, pp. 46-53.
- [7] N. Ahmed, A. L. C. Barczak, T. Susnjak, and M. A. Rashid, "A comprehensive performance analysis of Apache Hadoop and Apache Spark for large scale data sets using HiBench," *Journal of Big Data*, vol. 7, no. 1, Art. no. 1, Dec. 2020.
- [8] Y. K. Gupta and S. Kumari, "Performance Evaluation of Distributed Machine Learning for Cardiovascular Disease Prediction in Spark," *5th International Conference on Trends in Electronics and Informatics (ICOEI)*, 2021, pp. 1506–1512.
- [9] G.-M. Park, Y. S. Heo, and H.-Y. Kwon, "Trade-Off Analysis Between Parallelism and Accuracy of SLIC on Apache Spark," *IEEE International Conference on Big Data and Smart Computing (BigComp)*, 2021, pp. 5–12.
- [10] M. Haggag, M. M. Tantawy, and M. M. El-Soudani, "Implementing a deep learning model for intrusion detection on apache spark platform," *IEEE Access*, vol. 8, 2020, pp. 163660–163672.
- [11] H. Sayed, M. A. Abdel-Fattah, and S. Kholief, "Predicting Potential Banking Customer Churn using Apache Spark ML and MLlib Packages: A Comparative Study," *International Journal of Advanced Computer Science and Applications*, vol. 9, no. 11, 2018.
- [12] A. M. Mickaelian, "Big Data in Astronomy: Surveys, Catalogs, Databases and Archives," *Communications of the Byurakan Astrophysical Observatory*, vol. 67, pp. 159–180, 2020.
- [13] "HiBench." <https://github.com/Intel-bigdata/HiBench>.
- [14] M. Zaharia et al., "Apache spark: a unified engine for big data processing," *Communications of the ACM*, vol. 59, no. 11, Art. no. 11, 2016.
- [15] D. Borthakur, "HDFS architecture," Document on Hadoop Wiki. URL <http://hadoop.apache.org/common/docs/r0>, vol. 20, 2010.
- [16] "Apache Ignite." <https://ignite.apache.org/>.
- [17] C. Stan, A. Pandelica, V. Zamfir, R. Stan, and C. Negru, "Apache Spark and Apache Ignite Performance Analysis," *22nd International Conference on Control Systems and Computer Science (CSCS)*, May 2019, pp. 726–733.
- [18] P. Zhou, Z. Ruan, Z. Fang, M. Shand, D. Roazen, and J. Cong, "Doppio: I/O-aware performance analysis, modeling and optimization for in-memory computing framework," *IEEE International Symposium on Performance Analysis of Systems and Software (ISPASS)*, 2018, pp. 22–32.
- [19] "RubiX." <https://github.com/qubole/rubix>.
- [20] "Azure HDInsight." <https://docs.microsoft.com/en-us/azure/hdinsight/spark/apache-spark-improve-performance-iocache>.
- [21] "Databricks Delta Lake." [Online]. Available: <https://docs.databricks.com/delta/optimizations/delta-cache.html>.
- [22] "Open Cache Acceleration." <https://open-cas.github.io/>.
- [23] "Alluxio." <https://www.alluxio.io/>.
- [24] J. Pérez-Ortega, N. N. Almanza-Ortega, A. Vega-Villalobos, R. Pazos-Rangel, C. Zavala-Díaz, and A. Martínez-Rebollar, "The k-means algorithm evolution," *Introduction to Data Science and Machine Learning*, IntechOpen, 2019.
- [25] D. F. Gleich, "PageRank beyond the web," *Siam Review*, vol. 57, no. 3, Art. no. 3, 2015.
- [26] J. Liu, X. Li, and J. Dong, "A survey on network node ranking algorithms: Representative methods, extensions, and applications," *Science China Technological Sciences*, vol. 64, no. 3, Art. no. 3, 2021.
- [27] S. Jamil, T. Mohd, S. Masrom, and N. Ab Rahim, "Machine Learning Price Prediction on Green Building Prices," *IEEE Symposium on Industrial Electronics Applications (ISIEA)*, 2020, pp. 1–6.
- [28] N. H. Zulkifley, S. A. rahman, N. H. Ubaidullah, and I. Ibrahim, "House Price Prediction using a Machine Learning Model: A Survey of Literature," *International Journal of Modern Education and Computer Science*, vol. 12, 2020, pp. 46–54.
- [29] Daniilak, *Russia Real Estate 2018-2021*. 2021. [Online]. Available: <https://www.kaggle.com/mrdaniilak/russia-real-estate-2018-2021>.
- [30] C. Krauss, X. A. Do, and N. Huck, "Deep neural networks, gradient-boosted trees, random forests: Statistical arbitrage on the S&P 500," *European Journal of Operational Research*, vol. 259, no. 2, Jun. 2017, pp. 689–702.
- [31] A. D. Linder and R. D. Wolfinger, "Forecasting with gradient boosted trees: augmentation, tuning, and cross-validation strategies: Winning solution to the M5 Uncertainty competition," *International Journal of Forecasting*, 2022.

Structural Equation Modelling for Validating Disruptive Factors in Livestock Supply Chain

Nur Amly Abd Majid^{1*}, Noraidah Sahari², Nur Fazidah Elias³
Hazura Mohamed⁴, Latifah Abd Latib⁵, Khairul Firdaus Ne'matullah⁶

Computing Department, Universiti Selangor, Malaysia¹

Center for Software Technology and Management, Faculty of Information Science and Technology^{2,3,4}
Universiti Kebangsaan Malaysia, Malaysia^{2,3,4}

Communication Department, Universiti Selangor, Malaysia⁵

Centre for Foundation and General Studies, Universiti Selangor, Malaysia⁶

Abstract—The purpose of this paper is to deploy a structural equation modeling approach through the Partial Small Square technique to validate the disruptions factors that affect livestock supply chain performance. The disruption prediction factors were obtained from the analysis of literature studies and data from the Department of Veterinary Services (DSV) and expert evaluation. Factors considered in the study model are Livestock Process, Finance, Breeders, Quality, Facilities, Technology, Demand, Supply, Information Communication, Sales, Transportation, Government Involvement, Disaster and Syariah Compliance. The results of the study found that the factors of Livestock Process, Finance, Breeders, Livestock Quality, Technology, Supply, Sales, Transportation, Government Involvement and Syariah Compliance were accepted as disruptions in the livestock supply chain. The findings of this study will assist farmers and livestock stakeholders to take necessary measures to minimise the disruption and further the government's goal of enlivening small and medium livestock enterprises in Malaysia.

Keywords—Supply chain management; disruption; livestock; structural equation modelling

I. INTRODUCTION

Livestock breeding is an agricultural activity that is one of the most important in the Malaysian agricultural industry [1]. According to United Nations [2], the importance of animal husbandry is seen as the “intensification of animal production as a way to ensure food supply”. The increasing demand for livestock meat supply in developing countries, including Malaysia, is driven by population growth, urbanization, industry, and increasing income. Households, the influx of foreign workers and an increase in the tourism industry [3]. Based on the Department of Statistics Malaysia (JPM), livestock in Malaysia contributed 14.9 percent to Gross Domestic Product (GDP) in 2018. This situation shows the livestock industry is one of the important sub-sectors in agricultural development in the country.

As reported by Institute of Supply Chain Management [4], a supply chain is defined as the smooth design of management, where the value-added process flows smoothly across organisational boundaries to fulfil the real needs of end customers. A supply chain, is an integration activity that occurs between one network and another to obtain raw materials,

transform raw materials into a semi-final product form, and finally become a final product, and then deliver the final product to customers through the distribution system [5].

Fig. 1 is the supply chain of the livestock industry as a whole. The livestock supply chain is generally the same for all livestock animals starting from the feed supplier to the end-user and through different livestock processes [6].

Breeding data is necessary to ensure that breeders have access to the most up-to-date information and can make informed decisions during the breeding process. Preliminary research indicated that farmers have no experience using information systems in livestock management, which is one of the types of disruption. The Department of Veterinary Services has built a system of information and technology exchange channels based on previous studies to ensure information is conveyed swiftly and accurately. Its purpose is to bridge the gap between officers and farmers in terms of communication. However, the study indicated that, with the exception of family support, social interactions, and internal drive, information and communication technology elements have no positive and substantial impact on farmer success [7]. The study also discovered that breeder information systems designed to aid livestock sector stakeholders were underutilized [8]. As a result, the Department of Veterinary Services (DSV) finds it difficult to get data on breeders and animals for data analysis, and the process of recording livestock reports cannot be implemented on time due to the difficulty in obtaining livestock data.

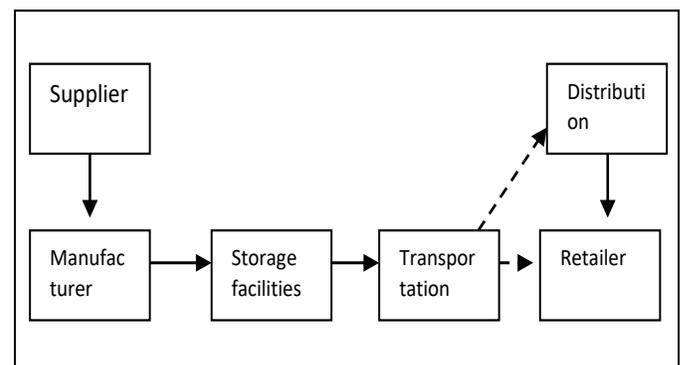


Fig. 1. Livestock Supply Chain.

*Corresponding Author.

The livestock industry in Malaysia is currently experiencing difficulties and disruptions in managing farms in a systematic manner to produce large-scale meat supplies, particularly to meet the demand of local consumers. Due to poor breeding rates, the cattle industry is experiencing a livestock population shortfall. Unsuitable breeds, high feed costs, and a lack of grazing land all contribute to the disruption. The role of private-sector involvement is also limited due to low-profit returns in the livestock sector, and this has resulted in the involvement of SME breeders still at a moderate level.

In addition, livestock practitioners in Malaysia often experience technical problems such as livestock selection, livestock breed selection, selection and provision of good livestock feed as well as farm management in accordance with the standards set by the government. The situation of livestock practitioners who still breed traditionally leads to unforeseen organizational management problems such as issues of destruction of livestock, management of labour or contract workers, credit loans and repayment methods that cannot be well followed by breeders. The issue of lack of DSV officers in some farming areas is seen to affect efforts to guide and channel accurate, up-to-date and direct information to farmers. Non-compliance of officials involved in financial management as well as lack of skills training also resulted in losses and tarnished the image of the government.

The demand for livestock meat supply in Malaysia is constantly increasing from year to year but the supply of stocks is always insufficient causing Malaysia to rely on import sources. The findings found that in 2017, a total of 167,439.17 metric tons of large ruminant meat were imported into the country [9]. The year 2020 saw the global Covid-19 pandemic had disrupted most of the country's economic activities due to movement control measures, including import sources of meat supply. To address this disruption, DSV is implementing efforts by providing guidance and educational activities to farmers as an incentive to increase meat production. However, the country still imports meat supplies from abroad. Table I refers to the time series statistics of the number of ruminants and pigs issued by the Veterinary Department from 2016 to 2020 in Malaysia.

TABLE I. NUMBER OF LIVESTOCK STATISTICS

Types of Livestock	2016	2017	2018	2019	2020*
Buffalo	59,740	54,632	48,195	47,652	47,266
Cow	654,602	620,521	589,113	581,567	585,597
Goat	350,370	318,032	289,361	256,159	264,922
Sheep	134,057	126,161	122,205	117,921	117,526
Pig	1,370,763	1,412,737	1,448,122	1,468,788	1,463,000
Chicken	-	-	-	138,689,878	145,049,672
Duck	-	-	-	6,971,200	6,481,782

II. LITERATURE REVIEW

Every organization faces supply chain disruption at some point during the manufacturing process. The supply chain is a system consisting of organizations, human resources, technology, activities, information and resources involved in the activity of converting raw materials into finished products and delivered to consumers. Each supply chain activity has a different purpose. The supply chain in each organization is different and depends on the activities as well as the end product produced. Disruptions that occur in the supply chain complicate many parties including suppliers and end-users and expose various risks and disadvantages to all stakeholders in the supply chain [10].

In recent years, the supply chain process has become longer and more complex while the level and frequency of supply chain disruptions have shown an increase [11]. Disruption is defined as an event that disrupts the flow of material in the supply chain that causes the movement of goods to stop abruptly [12]. Disruption occurs due to several factors such as natural disasters, labour disputes, dependence on a single supplier, suppliers experiencing bankruptcy, violence, war and political instability. Transportation accidents, failure of public places and product failure and disruption of disasters after the occurrence of disasters such as haze, water crisis, forest fires and others. Particularly in Malaysia, natural disasters and extreme weather conditions are to some extent a threat to supply chain [13]. According to Pfohl, et al. [11], apart from natural disaster disruptions, other disruption factors identified are production machine failure, quality problems on final products, quality problems at the resource level, system failures, and problems from human resources, suppliers' delays, delays transportation and natural disasters themselves.

Each supply chain is exposed to its own dangers as a result of increased supply chain transit [14]. Each phase of the chain becomes more susceptible to different types of interference as the chain process becomes more efficient. Through the literature review, disruption caused by inefficiency in management, in turn, disrupts organisational activities as well as increases the cost of operating expenses and repairs [15]. A survey conducted on 559 companies representing 62 countries and 14 various the industry sector found that 85 per cent of the companies reported at least one supply chain disruption that occurred in the past 12 months [16]. While a recent study conducted in 2019 by the Business Continuity Institute reported despite the awareness of supply chain risk has increased, most companies remain at high risk of being exposed to disruptions. The study found that 74 per cent of respondents from a survey of 426 organisations had experienced at least one disruption in the supply chain of which 6 to 20 disruptions were reported each year, which is 50 per cent of companies, and experienced financial losses varying from 50 thousand to 500 million euros. While more than 23% of businesses reported losses of at least one million euros as a result of disruptions. Supply chain disruptions not only cause financial loss but potentially damage a company's brand or reputation as a result of third-party failures. According to the study, 27 per cent of businesses suffered a tarnished reputation, 58 per cent lost productivity, and 38 per cent lost

revenue. With a damaged reputation of 11.6 per cent, Asian countries are in fourth place [17].

There's no denying that the Covid-19 pandemic has shifted the country's economic landscape in unexpected ways. The livestock industry is no exception. The pandemic disruption that has hit the world presents a new form of challenge to farmers in Malaysia who are struggling to get the desired results. Farmers and livestock stakeholders face uncertain income and other disruptions in the supply chain and in turn expect assistance from the government for extensive assistance plans and long-term efforts to ensure the welfare of farmers is protected. The impact of the Covid 19 pandemic resulted in the distribution chain being affected due to the closure of operations, the absence of employees and declining cash reserves [18]. Physically disrupted supply chain disruptions have prompted entrepreneurs to take alternative measures by switching to online sales through social media including Facebook and WhatsApp, product delivery through private drivers and downsizing businesses to save operating costs [19].

Referring to the Table II are the sources of research findings past a discussion of the disruption factors identified in the supply chain.

TABLE II. FINDINGS OF POST DISRUPTION STUDIES IN THE SUPPLY CHAIN

Disruption Factors	References Study Findings
Production Facility Failure	[20] [21] [5] [21] [25] [26]
Quality Problems at The Resource Level	[21] [23] [24] [25] [20] [26] [27] [28]
Information Technology System Failure	[20] [21] [30] [29]
Human Resource Issues	[22] [30] [29] [27]
Distribution Network Discontinued	[22] [30] [29] [27]
Demand Fluctuations	[27] [25] [28]
Supplier Delay	[27] [28]
Bankrupt Supplier	[20] [5]
Top Suppliers Bankrupt Suppliers	[22] [30] [29] [27]
Transportation Breakoff	[21] [27] [30] [29] [27]
Port Party Strike	[22] [30] [29] [27]
Natural Disasters	[23] [22] [30] [29] [27]
Security Risk (Terrorist Threat)	[22] [30] [29] [27]
Communication Failure	[30] [29] [27] [22]
Political and Economic Instability	[21] [22] [24][77]
Regulatory and Legal Risks	[20] [25] [21] [23] [22][73]

The study also considered the SCOR Model and PESTLE analysis and the Behzad disruption framework to identify disruption factors according to the livestock industry.

1) *Model SCOR*: The SCOR model was developed in 1996 by Pittiglio Rabin Todd and McGrath and endorsed by

the Supply Chain Management Council as an industry-standard model in supply chain management [31]. Referring to Fig. 2, SCOR is process-oriented consisting of Plan (planning), Source (source), Make (manufacturing), Deliver (delivery) and Return (return) as in Fig. 2. This process encompasses the entire supply chain process from the point of view of suppliers, organizations and customers. The organizations involved are internal and external organizations.

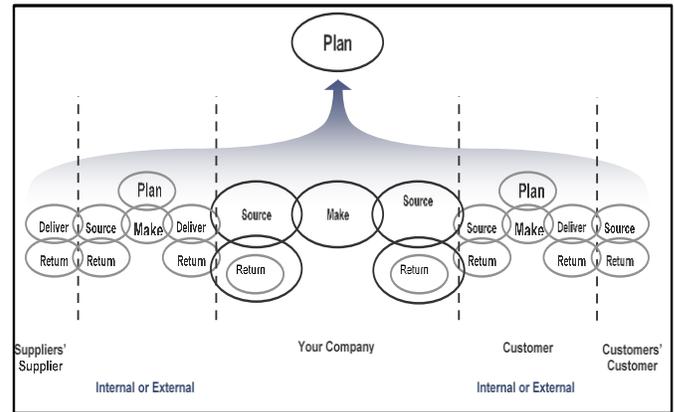


Fig. 2. SCOR Model.

2) *PESTLE Analysis*: There are six factors in the PESTLE framework. The factors are political, economic, social, technology, legal and environmental. PESTLE analysis is a method of analyzing the external environment to identify factors that contribute to the success or failure of an organization or industry. PESTLE analysis has a different structure from previous studies such as PEST and STEPE [32].

- **Political factors:** Politics plays an important role in business. In the livestock industry, politics plays a role in the effort to grow the industry through government involvement in assisting farmers by aiding further expand the livestock industry. For example, assistance from Majlis Amanah Rakyat to invest in industrial and commercial programs in the agriculture and food-based livestock sector is seriously and comprehensively needed to ensure that farmers rise up in the supply chain as well as increase the participation of Muslims in controlling the country's food supply meat industry.
- **Economic Factors:** Economic factors are measuring measures that are used to assess an organization's financial success. It is common that economic conditions often change over the life of an organization through comparisons of current levels of inflation, unemployment, economic growth and international trade. In the breeding process, financial difficulties are a huge and crucial issue. Farmers frequently complain about a lack of capital, which prevents them from increasing the number of farm animals. Only enough profit is made to be used as working capital.

- **Social Factors:** In a given situation or problem, social factors are used to assess the mentality of an individual or user. Demographic variables are also known as social factors. Indicators of social measurements are such as age distribution, population growth rate, employment level, income statistics, education level, religion, culture and social interaction. Apart from the aforementioned factors, the measurement of social factors also takes into account health and environmental concerns. At both the national and global levels, a variety of social and communication factors play a critical role. Among the several social sub-factors that can be considered to determine the measurement of social performance for an organization are social mobility, ethics and religion, lifestyle, level of education, historical issues, identity and beliefs, demographics and two-way cultural communication. The breeding process is influenced by the involvement of breeders who are 45 years old and older on average. The level of acceptance and openness in the breeding process makes the breeding field either a failure or a success and it results in a loss of cost and no return on capital that has been issued or vice versa. Similarly, the level of low or higher education among farmers affects the level of knowledge of the latest technology, and this can result in a lack of production of livestock products in the country or vice versa.
 - **Technology Factors:** When it comes to accurately assessing organisational performance, technology plays a big role. Technology advancements can improve internal efficiencies and prevent products or services from becoming obsolete. Every year the role of technology in the industry is increasing and technology in every chain is increasingly required to keep the process running effectively. Barn housing technology is one of the innovations that have been introduced into the livestock process [33]. Among other findings, livestock owners at present do not have to worry about the health of livestock because there is livestock automation technology based on the Internet of Things (IoT) that can monitor livestock remotely using drones and wireless network technology and able to collect data from sensors installed on livestock as well as water quality sensors in several water sources around livestock areas [34]. Among the technology sub-factors that can be considered to determine the measurement of technology performance for an organization are an information management system, quality and price, information change rate, minimize information retrieval problems, intellectual property, outsourcing, network coverage, patents and licenses, research and development, production efficiency and government legislative activities.
 - **Environmental Factors:** Environmental factors today are often seen as a threat to the environment. The livestock industry is no exception to environmental issues. Livestock activities, especially cattle breeding is carried out on the oil palm and rubber plantation lands that offer the potential to be cultivated in an integrated manner [9]. However, there are cases of livestock deaths recorded due to wild animal attacks, poisoning, negligence and deaths due to floods as well as accidents.
 - **Legal Factors:** The legal element is the final component of the PESTLE analysis. In this factor, the knowledge of laws and regulations needs to be known and learned importantly to prevent the occurrence of unnecessary legal costs. In carrying out livestock practices, there are guidelines that need to be followed through the Livestock Farm Practice Scheme to ensure good livestock practices to ensure the production of quality livestock and safe to eat. Good Livestock Practices (GAHP) MS 2027: 2006 includes livestock health management programs, biosecurity programs, sanitary and phytosanitary programs and farm waste and pollution management programs [9]. These farming practices cover all types of ruminant and non-ruminant livestock.
- Finally, PESTLE analysis is utilised to evaluate external factors that have an impact on an industry. The PESTLE analysis is an excellent starting point for developing the study framework [35]. Organizational owners need to identify the risks to be faced and use all of these factors and knowledge to make decisions to improve organisational performance. PESTLE can comprehensively understand the environmental picture of an organisation and can maximise opportunities as well as reduce the threat of disruption to the organisation [36].
- Political factors can determine how the direction of political parties affects business development and growth in animal husbandry. Economic factors are used to assess the impact of interest rates, taxes, stock markets, consumer confidence and other economic metrics. Breeders and livestock stakeholders need to be more competitive in line with current changes to withstand the challenges of lifting the livestock industry towards high-income economic transformation. Social factors affect lifestyle changes, advertising targets, ethics, demographics and culture. Technology factors are seen to help industry performance and ensure organizations use the latest technology in business. New high-tech approaches in the production of the livestock industry can help double the revenue from the livestock industry. Legal factors are expected to help regulate new laws and regulations that affect the operation of the industry and environmental factors can identify accidents and weaknesses that occur as well as solutions that can be considered, especially in the livestock industry.

3) *Behzad disruption framework*: Referring to Table III, the Behzad disruption framework is divided into three parts namely the organizational level, the network level consisting of demand, supply and transport and the environment level.

At the organizational level, there are several disorders identified like production machine failure, the occurrence of quality problems on the final product, the failure of information technology systems and the occurrence of problems from human resources due to the strike. While environmental disruptions are broken down into demand, supply, and transportation sub-factors. At the network level, disruptions at the demand level stem from the distribution network being stalled in the chain due to disruptions occurring in one of the chains and demand fluctuations in livestock supply. Disruption at the supply level is due to quality problems at the source level and the occurrence of some problems from the suppliers like delays and suppliers who suddenly experience losses (bankruptcy). Disruptions at the transportation level were identified as being caused by suppliers experiencing bankruptcy situations, transportation delays and strikes from employees. Meanwhile, disruption at the environmental level is caused by natural disasters, security risks like threats from terrorists, communication failures between several chains, political and economic instability and regulatory and legal risks. Findings from the study of [11] also found disruption factors also stem from natural disasters, security risks such as terrorist threats, the occurrence of communication failures between several chains, political and economic instability and regulatory and legal risks occurring at the chain network-level causing chain movements to pause.

TABLE III. BEHZAD FRAMEWORK

	Disruption Factors	
Organizational Level		Production machine failure
		Quality problems with the final product
		Information technology system failure
		Problems from human resources (strike)
Network Level	Demand	Network distribution disruption
		Demand fluctuates
	Supply	Quality problems at the resource level
		Supplier delays
		Supplier incurs losses (bankruptcy)
	Transportation	Bankruptcy of third-party logistics
		Transportation delays
		Strike
Environmental Level		Natural disaster
		Security risk (terrorist threat)
		Communication failure between several chains
		Political and economic instability
		Regulatory and legal risks

In total, this study presents 18 constructs of livestock disruption in the first circulation as in Table IV.

TABLE IV. LIST OF LIVESTOCK DISRUPTION CONSTRUCTS

No.		Factors	References
1	Organizational Level	Operating Process (Management)	[31][32] [37]
2		ICT Management Failure	[32][71]
3		Information Communication Disruption	[11] [38]
4		Livestock Process	[9] [39] [40] 41] [42] [43]
5		Problems of human development (Breeders)	[31][11][72]
6		Farm worker	[38] [9] [11] [39] [41] [42] [43]
7		Quality	[11][38]
8	Network Level	Supplies	[31][32][38]
9		Sales	[32]
10		Financial Assistance (External)	[11][37]
11		Finance (Internal)	[11][37][71][75]
12		Facilities (Facilities)	[31] [9] [42] [43]
13		Request	[11][72]
14	Environmental Level	Transportation	[11][37][38][31]
15		Flexible	[9][44] [46] [11] [45]
16		Natural disaster	[11][38]
17		Government Involvement	[32][38][76]
18		Security	[11][38][37]

III. METHODOLOGY

This study adopts the study design proposed by Marakas [47]. The research methodology includes six main phases namely the problem analysis phase, the initial study phase, the model development phase, the instrument development phase, the model validation phase and lastly the model feasibility phase. The evaluation in each phase is using the Mini Delphi method for the rounds that are required as in the Model Development Phase. Conforming to the Reffi et al. [73] Mini Delphi is a technique that uses a discussion-based approach between moderators and involved experts. This study used a four-round Mini Delphi approach formally through face-to-face and email methods. In accord with Azizah et al. [74] panel responses were analyzed to identify the mean value for each construct.

The first phase of our research was to identify concerns and questions. Then the objective of the study was identified based on the issues and questions of the study, background and previous studies. This step of research results in a conceptual disruptor. The second phase includes a survey questionnaire as well as interviews with experts and stakeholders. Other factors

that lead to disruption other than those listed in the literature review are identified at this stage. Third phase: Assessment and selection of disruptive variables using a checklist tool. Data gathering based on the specified questionnaire in the fourth step. Model validation based on statistical analysis is the fifth phase. The Partial Small Square method was used to assess validation using the Structured Equation Modelling approach. The sixth phase entails creating a prototype of an information system based on the validated model.

IV. ANALYSIS

The data was analyzed using the Structured Equation Modelling (SEM) through Partial Least Squares Method. Structural Equation Modelling is a second-generation data multivariate analysis method used to test linear theory and causal augmentation models [48],[49],[50]. Analysis through Partial Least Squares Method approach was implemented through two levels of analysis. The first step involves examining the validity and reliability of the measurement model while the second step involves the evaluation and interpretation of the structural (theoretical) model [51]. The following is an explanation related to the analysis.

A. Convergence Validity

Convergent validity is defined as the degree to which some indicator can measure a given concept [50]. [52] proposed several criteria to measure the validity of convergence, which are factor loading, Cronbach Alpha (CA), Composite Reliability (CR) and Mean Variable Extraction (AVE). He also suggested that the load factor of the items should be greater than 0.7. The second criterion for convergent validity to convergence is composite reliability which refers to the degree to which a set of items consistently measures latent variables [52].

Through analysis, the Cronbach Alpha value and composite reliability were checked. The Cronbach Alpha value ranged from 0.8 to 0.873 while the composite reliability ranged from 0.873 to 0.902, which is significantly higher than the recommended level of 0.7 [53],[52]. Therefore, this result confirms that the validity of the convergent model has been tested. In addition, the average variance extracted (AVE) values were corrected to confirm the convergent validity of the external model. AVE is the mean-variance taken from several items related to the variance shared by the measurement error. In other words, AVE measures the variance shared by the metrics against the measurement error. If the AVE value is at least 0.5, then a latent variable can be inferred [52]. In this study, the AVE values ranged from 0.511 to 0.728, indicating that the study design constructs were validated [52].

B. Discriminant Validity

Discriminant validity has been used to measure the extent to which the constructs in the model differ from one another [50]. This validity is important because it ensures that the constructs in the model are unique and do not have high affinity for each other. In other words, items that measure the proposed construction should have high load, while items that do not measure the proposed construction will have low load. In the study, the validity of discrimination was measured using three criteria: Fornell Lacker, Cross Loading and HTMT.

1) *Fornell lacker*: The result of the square root of the AVE for the model's structures is placed on the diagonal of the correlation matrix. The model will be declared discriminant if the value of the square root AVE of each structure is greater than the elements in the column and row of each structure. The results of the study showed that the validity of the discriminant was confirmed because the squared value of AVE for each structure was greater than the mutual correlation of the columns and rows of the structure.

2) *Cross loading*: Cross-loading is the second approach to measuring the validity of discrimination. In this method items that are matched to the proposed constructs will have a high factor load while items matched to the proposed constructs will have a low load.

3) *Heteroit Monotrait (HTMT)*: Heterotope-monomer correlation ratio (HTMT) was used to assess the validity of the discrimination considered to be more accurate than other methods [54]. HTMT is recommended because it achieves higher specificity and sensitivity compared to cross criteria. An HTMT value close to 1 indicates invalid discrimination. Some authors recommend a cutoff of 0.85 [55], while others suggest a value of 0.90 [56]. If the value of HTMT is greater than this threshold, then selectivity is not valid. Result shows the HTMT clearly indicating that the HTMT value is less than 0.90 and further validates the validity of the discrimination.

C. Structural Model Assessment

Structural Model Evaluation will be carried out once the validation model has been validated. Generally, there are several approaches to measuring the structural model of multicollinearity, R-square, relevant and predictive coefficient routes.

1) *Multicollinearity*: Multicollinearity exists when two or more exogenous variables have a very high correlation [57]. It shows that some exogenous variables can be explained by other exogenous variables. Multicollinearity may result in inflationary problems of standard regression coefficients, which results in significant reduction in inflation [58]. Multicollinearity is said to occur when the correlation coefficient value is greater than 0.90 [52]. Additionally, collinearity issues can also be examined with reference to VIF values and tolerance. The author in [59] states that the VIF value should not exceed 5 to confirm that the structural model has no multicollinearity problem. The Fig. 3 shows the results of the alignment statistic (inner VIF) for each element. Each factor falls within the mean value from 1.102 to 2.812, which is within an acceptable range of less than 5.

2) *R-Square*: R-square is a measure of the predictive accuracy of a model and it is also considered as the combined effect of exogenous variables on endogenousness [60]. In other words, R-square provides the number of variants of endogenous variables that can be described by endogenous variables. In the PLS-SEM model, the R-square coefficient values of 0.67, 0.33 and 0.19 are classified according to three force levels, respectively, as medium and low [61]. According

to Henseler et al. [62], when structural models are explained by one or two modest exogenous variables, the R-square value is acceptable, and if the endogenous latent variables depend on some exogenous variable, the value of R-square is acceptable. The Fig. 4 shows the R-square value has significant level.

3) *F Square*: The F-square assesses the relative impact of each predictor construct on endogenous constructs [63]. Specifically, it measures the strength of an exogenous variable that impacts the endogenous variable in the R-square. According to the guidelines developed by [64], F-square values 0.02, 0.15 and 0.35 are considered as small, medium and strong [65]. The Fig. 5 shows the effect of the size of each exogenous variable on the endogenous variable.

Factor	Disruption
Transport	1.429
Supply	2.812
Disaster	1.889
Sales	1.701
Infrastructure	1.157
Government	1.991
Financial	1.902
Communication	1.932
Quality	1.559
Breeder	1.927
Demand	2.749
Livestock process	1.579
Syariah	1.102
Technology	2.179

Fig. 3. Multicollinearity.

Construct	R Square	R Square Adjusted
Disruption	0.795	0.785

Fig. 4. R-square Test Results.

Factor	Disruption
Transport	0.027
Supply	0.088
Disaster	0.001
Sales	0.017
Infrastructure	0.007
Government	0.069
Financial	0.045
Communication	0.035
Quality	0.038
Breeder	0.067
Demand	0.022
Livestock process	0.098
Syariah	0.03
Technology	0.228

Fig. 5. F-Square Test Results.

4) *Predictive relevance or blindfolding*: Q-Square analysis was performed to measure the relevance of exogenous constructs in predicting endogenous constructs [66],[67],[50]. When the Q-square value is higher than zero, this means that the model has a prediction relation, and if the value is zero and below, it indicates a lack of predicted prediction [68]. Based on result, the value of Q2 obtained by 0.415 is greater than the value of 0 which means that some exogenous variables can predict endogenous variables.

5) *Route coefficient*: Route coefficient is the standard version of the linear regression used to assess whether the proposed hypothesis is statistically significant or not significant. Each hypothesis proposed by the model is determined whether it is significant by means of path coefficient analysis. PLS-SEM uses a 5000-sample bootstrapping approach for hypothesis testing. A 95% confidence level with alpha 0.05 was used for hypothesis testing. The hypotheses show that a p value less than 0.05 is significant while a p value greater than 0.05 is not significant.

The Table V analysis results showed that 10 out of 14 hypotheses showed a significant value of $p < 0.05$. Significant results mean that there is significant impact of exogenous constructs on endogenous constructs.

TABLE V. HYPOTHESIS RESULTS

Hypotheses	Relation	O. S.	S. M.	S. D.	Nilai t	p values
H ₁	Trasport -> Disruption	0.089	0.087	0.031	2.892	0.0040
H ₂	Supply -> Disruption	0.225	0.228	0.059	3.833	0.0000
H ₃	Disaster -> Disruption	0.021	0.024	0.039	0.538	0.5910
H ₄	Sales -> Disruption	0.077	0.075	0.038	2.01	0.0450
H ₅	Infrastructure -> Disruption	-0.04	0.039	0.025	1.607	0.1090
H ₆	Government-> Disruption	0.167	0.164	0.039	4.242	0.0000
H ₇	Financial -> Disruption	0.132	0.13	0.037	3.603	0.0000
H ₈	Communication -> Disruption	0.006	0.006	0.039	0.152	0.8790
H ₉	Quality -> Disruption	0.107	0.108	0.035	3.02	0.0030
H ₁₀	Breeder -> Disruption	0.122	0.12	0.036	3.418	0.0010
H ₁₁	Demand -> Disruption	0.006	0.008	0.046	0.126	0.9000
H ₁₂	Livestock process -> Disruption	0.147	0.148	0.035	4.201	0.0000
H ₁₃	Syariah -> Disruption	0.071	0.073	0.029	2.431	0.0150
H ₁₄	Technology -> Disruption	0.21	0.207	0.048	4.384	0.0000
<i>OS=Original Sample/SM=Sample Mean/SD=Standard Deviation</i>						

Thus, the ten constructs accepted in this study explain the disruptions in the livestock supply chain that occur and impact the livestock sector. The construct of this disruption needs to be emphasized so that the competitiveness and sustainability of the livestock sector can be enhanced from time to time.

V. CONCLUSION

Positive results are presented for H1, H2, H4, H6, H7, H9, H10, H12, H13 and H14 where the constructs of transport, supply, sales, government involvement, finance, quality, livestock, livestock processes, syariah and technology have a strong relationship positive to the disorder. The results obtained are in line with the disorder statement presented by Behzad [11] and Park et al. [69] which is found that disruptions factor consist of internal, external, supplier and end-user interference. The results of hypotheses H1, H2, H4, H6, H7, H9, H10, H12, H13 and H14 indicate disruptions occur and are accepted at the levels described by Park et al. [69]. This statement is also expressed by Gunasekaran et al. [38].

Meanwhile for H3, H5, H8 and H11 where the construct of disaster, facilities, communication and demand, the p value > 0.05. This indicates the construct did not have an impact as a disruption in the study. This result contradicts the Martha [12] statement that disruption can be caused by natural disasters, labour disputes, dependence on a sole supplier, suppliers experiencing bankruptcy situations, violence, war and political instability. Similarly, according to Fu et al. [70], communication in the supply chain is the latest advancement in information technology and scientific management that allows most industries to obtain and share information but communication is rejected as a distraction in this analysis. The demand construct was also rejected as a disruption in this study. The results of the demand hypothesis show that the supply-demand whether high or low has no role as a disruption although the study shows that the demand for livestock meat supply in Malaysia is constantly increasing from year to year and still unable meet the demand. This is because Malaysia has the option of relying on imported goods [9].

The future studies consider increasing the number of respondents, respondents with breeder status as well as expanding the sampling of the study to produce more accurate and comprehensive analytical results. In keeping with the next studies also need to consider external constructs like livestock management skills, livestock management experience, current planning and economics, social relationships, networking and marketing as well to be evaluated and given due attention in the livestock disruption model.

ACKNOWLEDGMENT

This work is supported by Universiti Kebangsaan Malaysia through the university research grant code GUP-2017-099 and FTM2.

REFERENCES

[1] Nik Muhamad Adnan, M.N. 2019. Pendidikan Pengembangan dan Amalan Penternak Lembu Pedaging Dalam Program Kawasan Tumpuan Sasaran Di Semenanjung Malaysia.

[2] Pertubuhan Bangsa-bangsa Bersatu. 2016. "The History of Factory Farming", United Nations.

[3] Norizan, R., Fariha, R., Dani, S. 2019. Urbanization and Quality of Life: A Comprehensive Literature. Journal Of Social Transformation And Regional Development Vol. 1 No. 2 (2019) 24-32.

[4] Institute of Supply Chain Management (2020).

[5] Sodhi, M.S. and Lee, S. 2017. An analysis of sources of risk in the consumer electronics industry, Journal of the Operational Research Society 58(11):1430-1439.

[6] RISDA 2018; Malaysia Rubber Industry Smallholders Development Authority.

[7] Alam, S.S., Jani, M.F.M. and Omar, N.A. (2011) An Empirical Study of Success Factors of Women Entrepreneurs in Southern Region in Malaysia. International Journal of Economics and Finance, 3, 166-175.

[8] Veronica Saiz-Rubio & Francisco Rovira Mas. 2020. From Smart Farming towards Agriculture 5.0: A Review on Crop Data Management; Agronomy 2020, 10, 207.

[9] Jabatan Perkhidmatan Veterinar.2018. Sistem Maklumat Veterinar Selangor (Vetsel).

[10] Pfohl, H. C., Kohler, H., & Thomas, D. 2018. State of the art in supply chain risk management research: empirical and conceptual findings and a roadmap for the implementation in practice. Logistics Research, 2(1): 33-44.

[11] Behzad, B., Adhitya, Arief, Lukszo, Zofia & Srinivasan, R. 2012. How to Handle Disruptions in Supply Chains – An Integrated Framework and a Review of Literature (July 20, 2012).

[12] Martha, C. W. 2018. The impact of transportation disruptions on supply chain performance, Transportation Research Part E 43 (2007) 295–320.

[13] Yusof, Rohana, S., Mohd Fo'ad, K., Mohamad Sukeri. 2010. Impak bencana banjir terhadap industri: Kes kajian di Kedah. In: International Seminar on Economic Regional Development, Law and Governance in Malaysia and Indonesia, 7-9 June 2010, Universitas Islam Riau Indonesia, Pekanbaru, Riau.

[14] Linghe, Y. and Abe, M. 2012. The impacts of natural disasters on global supply chains.

[15] Bogataj, D. and Bogataj, M. 2007. Measuring the supply chain risk and vulnerability in frequency space. International Journal of Production Economics, Vol. 108 Nos 1/2, pp.291-301.

[16] Business Continuity Institute (2011), Business Continuity Institute.

[17] Supply Chain Resilience Report. 2019. <https://www.thebci.org/uploads/assets/e5803f73e3d54d789efb2f983f25a64d/BCISupplyChainResilienceReportOctober2019SingleLow1.pdf>.

[18] Noor Fzlinda Fabeil, Khairul Hanim Pazim, Juliana Langgat. 2020. The Impact of Covid-19 Pandemic Crisis on Micro-Enterprises: Entrepreneurs' Perspective on Business Continuity and Recovery Strategy.

[19] Mahdi Jemmali, Loai Kayed B.Melhim and Mafawez Alharbi, "Multi-criteria Intelligent Algorithm for Supply Chain Management" International Journal of Advanced Computer Science and Applications (IJACSA), 10(4), 2019.

[20] Zsidisin, G.A. and Wagner, S.M. 2019. Do Perceptions Become Reality? The Moderating Role of Supply Chain Resiliency on Disruption Occurrence. Journal of Business Logistics, 31, 1-20.

[21] Manuj, I. and Mentzer, J.T. 2018. Global supply chain risk management strategies, International Journal of Physical Distribution and Logistics Management 38(3): 192-223.

[22] Stecke, K.E. and Kumar, S. 2019. Sources of supply chain disruptions, factors that breed vulnerability, and mitigating strategies, Journal of Marketing Channels 16(3): 193-226.

[23] Oke, A. and Gopalakrishnan, M. 2019. Managing disruptions in supply chains: A case study of a retail supply chain, International Journal of Production Economics 118(1): 168-174.

[24] Matook, S., Lasch, R. and Tamaschke, R. 2019. Supplier development with benchmarking as part of a comprehensive supplier risk management framework, International Journal of Operations and Production Management 29 (3): 241-67.

[25] Chopra, S. and Sodhi, M.S. 2018. Managing Risk to Avoid Supply-Chain Breakdown, Sloan Management Review 46(1): 53-61.

- [26] Roth, A.V., Tsay, A.A., Pullman, M.E. and Gray, J.V. 2008. Unraveling the food supply chain: Strategic insights from China and the 2007 recalls, *Journal of Supply Chain Management* 44(1): 22-39.
- [27] Blackhurst J., Rungtusanatham MJ., Scheibe K., & Ambulkar S., 2018. Supply chain vulnerability assessment: A network based visualization and clustering analysis approach. *J Purch Supply Manag.* in press.
- [28] Buscher & Wels .2010. The ISPS code and the cost of port compliance: An initial logistics and supply chain framework for port security assessment and management, *Maritime Economics and Logistics* 6(4): 322-348.
- [29] Faisal, M. N., Banwet, D. K. and Shankar, R. 2017. Information risks management in supply chains: an assessment and mitigation framework, *Journal of Enterprise Information Management* 20(6): 677-699.
- [30] Sutton, S. G. 2016. Extended-enterprise systems' impact on enterprise risk management, *Journal of Enterprise Information Management* 19(1): 97-114.
- [31] Huang, S.H., Sheoran, S.K. and Wang, G. 2016. A review and analysis of supply chain operations reference (SCOR) modell, *Supply Chain Management: An International Journal*, Vol. 9 No. 1, pp. 23-9.
- [32] Richardson, Jr. J. V. 2006. The library and information economy in Turkmenistan. *IFLA Journal*, 32(2), 131-139.
- [33] Nor Amna A'liah Mohammad Nor . 2018. Kajian Penilaian Tahap Penggunaan Teknologi Bagi Subsektor Pertanian Terpilih Ke Arah Pertanian Moden: Lembu Tenusu.
- [34] Rosdiadee. 2019. Projek LoRa (Long Range) by Prof Madya Ir Dr Rosdiadee Nordin; Pusat Kejuruteraan Elektronik dan KomunikasiT ermaju (PAKET), Fakulti Kejuruteraan dan Alam Bina (FKAB).https://www.ukm.my/news/Latest_News/teknologi-tanpa-wayar-melalui-dron-bantu-peladang-pantau-ternakan-haiwan-peliharaan/.
- [35] Hazem M Bani Abdoh1, Syarilla Iryani A. Saany2, Hamid H. Jebur3, Yousef A.Baker El-Ebiary. 2020. The Effect of PESTLE Factors on E-Government Adoption in Jordan: A Conceptual Model; October 2020; *International Journal of Engineering Trends and Technology*.
- [36] Alarussi & Alhaderi, 2018. Factors affecting profitability in Malaysia; June 2018; *Journal of Economic Studies* 45(1):00-00.
- [37] Thattapon Surasak, Nungnit Wattanavichean, Chakkrit Preuksakarn and Scott C.-H. Huang, "Thai Agriculture Products Traceability System using Blockchain and Internet of Things" *International Journal of Advanced Computer Science and Applications (IJACSA)*, 10(9), 2019.
- [38] Gunasekaran, A., Patel, C. and McGaughey, R.E. 2016. A framework for supply chain performance measurement, *International Journal of Production Economics*, vol. 87, no. 3, pp. 333-347.
- [39] Olutayo O, Babalobi. 2007. Veterinary geographic information systems applications in Nigeria: Limitations, challenges and needs. *Vet. Ital.*, 43(3): 491- 499.
- [40] Dani, S., & Deep, A. 2010. Fragile food supply chains- Reacting to risks, *International Journal of Logistics Research and Applications* 13(5): 395-410.
- [41] K. P. Sonavale, M. R. Shaikhand M. M. Kadamand V.G. Pokharkar. 2020. Livestock Sector in India:A Critical Analysis; *Asian Journal of Agricultural Extension, Economics & Sociology*38(1): 51-62, 2020; Article no. AJAEES.54091; ISSN: 2320-7027.
- [42] Pinak Ranade dan Asima Mishra .2015. Livestock Information Management System Web-GIS based (WGLIMS); *Int. Journal of Applied Sciences and Engineering Research*, Vol. 4, Issue 2, 2015.
- [43] T.Rohmawati and F.Ramadhani.2020. Android Based Livestock Sales Application Information System IOP Conf. Ser.: Mater. Sci. Eng. 879 012005.
- [44] Chan. 2015. Measuring the comparative advantage of agricultural activities: Domestic resource costs and the social cost-benefit ratio, *American Journal of Agricultural Economics*, 77, 243-250.
- [45] Christophe Feder, 2017. The effects of disruptive innovations on productivity. Article in *Technological Forecasting and Social Change*. May 2017.
- [46] Benita M. Beamon. 1999. Supply chain design and analysis : Models and methods. *Int. J. Production Economics* 55 (1998) 281-294.
- [47] Marakas, M.G. 1999. *Decision Support Systems in the First Century*. New Jersey: Prentice Hall.
- [48] Davcik, N.S. 2013. The use and misuse of Structural Equation Modeling in management research.1-40.
- [49] Wong, K.K.K. 2013. Partial Least Squares Structural Equation Modeling (PLS-SEM) Techniques Using SmartPLS. *Marketing Bulletin*, 2013, 24, Technical Note 1.
- [50] Hair Jr, J. F., Sarstedt, M., Ringle, C. M., & Gudergan, S. P. 2017. *Advanced issues in partial least squares structural equation modeling*. SAGE Publications.
- [51] Urbach, N., & Ahlemann, F. 2010. Structural Equation Modeling In Information Systems Research Using Partial Least Squares. *Journal Of Information Technology Theory And Application*, 11(2), 5-40.
- [52] Hair, J F., William C. Black, Barry J. Babin, and Rolph E. Anderson (2010), *Multivariate Data Analysis*, Englewood Cliffs, NJ: Prentice Hall.
- [53] Fornell, C., & Larcker, D. F. (1981). Structural equation models with unobservable variables and measurement error: Algebra and statistics. *Journal of marketing research*, 382-388.
- [54] Henseler, J., Ringle, C. M., & Sarstedt, M. (2015). A new criterion for assessing discriminant validity in variance-based structural equation modeling. *Journal of the academy of marketing science*, 43(1), 115-135.
- [55] Kline, R. B. (2011). *Convergence of structural equation modeling and multilevel modeling*. na.
- [56] Teo, T., Luan, W. S., & Sing, C. C. (2008). A cross-cultural examination of the intention to use technology between Singaporean and Malaysian pre-service teachers: an application of the Technology Acceptance Model (TAM). *Journal of Educational Technology & Society*, 11(4).
- [57] Gaur, A. S., & Gaur, S. S. (2006). *Statistical methods for practice and research: A guide to data analysis using SPSS*. Sage.
- [58] Gotz, O., Liehr-Gobbers, K., & Krafft, M. (2010). Evaluation of structural equation models using the partial least squares (PLS) approach. In *Handbook of partial least squares* (pp. 691-711). Springer, Berlin, Heidelberg.
- [59] Ramayah, T., & Lee, J. W. C. (2012). System characteristics, satisfaction and e-learning usage: a structural equation model (SEM). *Turkish Online Journal of Educational Technology-TOJET*, 11(2), 196-206.
- [60] Ramayah, T., & Scholtz, B. (2018, July). Role of Absorptive Capacity in Predicting Continuance Intention to Use Digital Libraries: An Empirical Study. In *Emerging Technologies in Computing: First International Conference, iCETiC 2018*, London, UK, August 23-24, 2018, *Proceedings* (Vol. 200, p. 297). Springer.
- [61] Chin, W. W. (1998). The partial least squares approach to structural equation modeling. *Modern methods for business research*, 295(2), 295-336.
- [62] Henseler, J., Ringle, C. M., & Sinkovics, R. R. (2009). The use of partial least squares path modeling in international marketing. In *New challenges to international marketing* (pp. 277-319). Emerald Group Publishing Limited.
- [63] Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* 2nd edn.
- [64] Cohen, J. (1992). Statistical power analysis. *Current directions in psychological science*, 1(3), 98-101.
- [65] Mohammad, J., Quoquab, F., Idris, F., Al-Jabari, M., Hussin, N., & Wishah, R. (2018). The relationship between Islamic work ethic and workplace outcome: A partial least squares approach. *Personnel Review*.
- [66] Stone, M., 1974. Cross-validatory choice and assessment of statistical predictions. *J. R. Stat. Soc. Ser. B Stat. Methodol.* 36 (2), 111-147.
- [67] Geisser, S. (1974). A predictive approach to the random effect model. *Biometrika*, 61(1), 101-107.
- [68] Henseler, J., Ringle, C. M., & Sinkovics, R. R. (2009). The use of partial least squares path modeling in international marketing. In *New challenges to international marketing* (pp. 277-319). Emerald Group Publishing Limited.
- [69] Park K., Min H. & Min S. 2016. Inter-relationship among risk taking propensity, supply chain security practices, and supply chain disruption occurrence. *Journal of Purchasing and Supply Management*, 22(2): 120-130.

- [70] Fu, W., & Menzies, T. 2017. Easy over hard: a case study on deep learning. Proceedings of the 2017 11th Joint Meeting on Foundations of Software Engineering - ESEC/FSE 2017.
- [71] Kok Yong Chan, Johari Abdullah and Adnan Shahid Khan, "A Framework for Traceable and Transparent Supply Chain Management for Agri-food Sector in Malaysia using Blockchain Technology" International Journal of Advanced Computer Science and Applications(IJACSA), 10(11), 2019.
- [72] Hicham Lamzaouek, Hicham Drissi and Naima El Haoud, "Digitization of Supply Chains as a Lever for Controlling Cash Flow Bullwhip: A Systematic Literature Review" International Journal of Advanced Computer Science and Applications (IJACSA), 12(2), 2021.
- [73] Mohd Reffi Hidayat Roslan, Kamsuriah Ahmad, Mohannad Moufeed Ayyash (2020); Factors Influencing Information Systems Quality From The System Developers Perspective. Asia-Pacific Journal of Information Technology and Multimedia Jurnal Teknologi Maklumat dan Multimedia Asia-Pasifik Vol. 9 No. 1, June 2020: 82 – 93.
- [74] Azizah Jaafar, Chan Siew Lee dan Nor Azan Mat Zin (2012) Pembangunan dan Penilaian Kepenggunaan Perisian Kursus Pendidikan Seksualiti Malaysia. Asia-Pacific Journal of Information Technology and Multimedia Jurnal Teknologi Maklumat dan Multimedia Asia-Pasifik Vol. 1 No. 1, June 2012 e-ISSN:2289-2192.
- [75] Hooman Shababi Haniyeh Ziari (2021). Explaining The Impact Of Information Technology Capabilities And Organizational Acceleration On Performance Of Banking Sector; Asia-Pacific Journal Of Information Technology And Multimedia Jurnal Teknologi Maklumat Dan Multimedia Asia-Pasifik Vol. 10 No. 2, December 2021: 37 - 50 E-Issn: 2289-2192.
- [76] Ismail, M. A., Nurul Afifah Abdul Alim, N. A., Mohd Shafiai, M. H., & Mawar, M. Y. (2022). Impak Harga Minyak Sawit terhadap Pertumbuhan Sektor Pertanian. Jurnal Ekonomi Malaysia, 56(1),
- [77] Leonard Lam, F. L., , Law, S. H., Azman-Saini, W. N. W., Khair-Afham, M. S. M., & Goh, L. T. (2022). High Technology Trade, Innovation and Economic Growth: Evidence from Aggregate and Disaggregate Trade Products. Jurnal Ekonomi Malaysia, 56(1).

IoT Enabled Smart Parking System for Improvising the Prediction Availability of the Parking Space

Anchal, Pooja Mittal

MDU, Department of Computer Science, Rohtak, India

Abstract—Smart cities are a result of persistent technological advancements aimed at improving the quality of life for their residents. IoT-enabled smart parking is one of the foundations of smart transportation which seeks to be versatile, long-lasting, and integrated into a Smart City. One of the studies shows that the drivers who are searching for free parking space can cause congestion problems up to 30%. There is a possibility to reduce air pollution and fluidity noise traffic by combining Internet of Things (IoT) sensors positioned in different parking areas with a mobile application and help the drivers to search for free places in different areas of the city and also provide guidance toward the parking space. In this paper, we show and explain a unique Data Mining-based Ensemble technique for anticipating parking lot occupancy to reduce parking search time and improve car flow in congested locations, with a favorable overall impact on traffic in urban centers. In this paper multi scanning, IoT Enabled smart parking model is proposed along with ensemble classifier that improvises the predictive availability of the free parking space. The predictors' parameters were additionally optimized using a Bootstrap and bagging algorithm. The proposed method was tested an IoT dataset containing a number of sensor recordings. The tests conducted on the data set resulted in an average mean absolute error of 0.07% using the Bagging Regression method (BRM).

Keywords—IoT; Data mining; sensors; ensemble; decision tree; bagging technique; boosting technique

I. INTRODUCTION

Data Mining plays a significant role in the digitalization of the technologies fusion with the Internet of things, virtual reality, 5G connectivity, and many more. This technology's influence is not only restricted to information systems [1]. It also affect other sectors too such as design, transportation, healthcare, shipping, and business procedures. The integration and amalgamation of IoT, Data Mining, and pervasive interconnection can accumulate enormous data inside a city, distribute the data to the central database, and use it for processing. For example, to make city parking and operations more efficient, as well as to support the household waste and hospitality sectors [2]. It is important to recognize that a "Smart City" is not only a city that is interconnected with IoT technology or simply "uses" the data connection. Future Smart City is an interconnected system, a "brain," in which innovation, management, infrastructure, and residents all communicate with one another [3]. IoT and Data mining allows us to track the patterns in a city's everyday existence, then modify them according to the requirements to create them more effective and real. Tragedies, accidents for example, is discovered by direct reporting, permitting for quicker action

and actual traffic reporting to suggest alternate paths to escape congestion bottlenecks, simplifying movement and lowering consumption and CO2 emissions. In addition, to the persistent parking issue in large cities, Data Mining can assist both in terms of detecting and reporting to residents the available parking spots in the neighborhood, as well as in terms of emphasizing to governments where new ones are needed.

According to Juniper Research's "Smart Cities: Leading Platforms, Segment Analysis & Forecasts 2019–2023," innovative solutions for smart city traffic, which are used to alleviate traffic jams in towns, would produce \$4.4 billion in revenue in 2023, up from \$2 billion in 2019. Smart parking methods, which are a result of technological advancement, are causing a revolution in the parking industry. Nowadays, finding a parking space in a city is a time-consuming and difficult process that can quickly turn into a nightmare. Due to the Internet of Things and Data Mining, it become possible to examine, and discover parking utilizing technology instruments like sensors, cameras, linked, and integrated things. As a result, smart parking becomes an important component of smart cities. IoT enabled Smart parking is an approach that integrates technology and human to use fewer resources (time, energy, and area) for attaining quicker, and optimum parking of the vehicles during the time when the cars are not in use[4]. To put it another way, IoT-enabled smart parking systems determine which parking areas are engaged and which are vacant, as well as build a real-time parking map. Real-time maps are useful in different areas.

- Assist the drivers to search the vacant space rapidly by using the mobile application.
- Ability to make actual information accessible, which allows the agents to spot any problems.
- Assist individuals in deciding on another mode of transport if the parking area is unavailable.

To conclude, IoT enabled smart parking system is made up of sensors, collection of actual records and analytics, and a smart payment methods which permit users to search for parking space in the preferred location and make the payment in advance if required. Traffic is severely limiting mobility in cities around the world, resulting in enormous expenses not only for travelers but for society itself. According to the Texas Transportation Institute, the average U.S. passenger spent 34 hours in traffic in 2010, up from 14 hours in 1982, and is likely to spend more than 40 hours in 2020. In areas where people work and reside, traffic contributes to pollution and noise. IoT-enabled smart parking provides a solution to the problem.

As the above scenario is described, to designing a predictive analytics framework by using the predicting algorithm and IoT-enabled sensor-generated data is a big challenge. It can also be used in a variety of real-world situations, such as air pollution forecasts, novel healthcare services, weather forecasts, etc. [5, 6, 7]. We propose a predictive technique in the Smart Mobility realm, with an emphasis on parking space availability. To solve the cities' parking problem, we are providing a consistent immediate estimation of parking space availability. In this regard, the goal is to develop a reliable structure of day-to-day 1-hour and 4-hour range estimations based on data provided by IoT sensors. Data mining techniques are used to produce a 1/2-hours-horizon prediction of parking space availability, improving the prediction of the well-known methodologies.

A. Organization of the Article is as Follow

The paper is organized as follows. Literature survey is presented in Section 2. In Section 3 methodology is described that contains a block diagram of multi-scanning model for IoT Enabled smart parking system. In this section, an overview of different techniques are also described that will help in the prediction of parking space. In Section 4 the performance of these techniques is represented using the performance measure R2 and Mean absolute error (MAE). And finally, conclusion is presented in Section 5.

II. RELATED WORK

The section is related to the review of the literature and seeks to highlight the key terms, perceptions, and flows of thinking in the area of parking through the readings. The important concern is the Parking; the following section emphasizes the features relevant to the precise objectives of the study.

There are two subsystems in the parking management system. A vehicular detection method (VDM) and a vehicular control method (VCM) are both available. The VDM monitors the availability of free spaces for parking and delivers the data to the VCM module for distribution of the availability of the space to drivers [8]. An intelligent parking system is also proposed by another author Kumar et al.[9]. The authors compared different types of sensors used in smart parking systems and data generated by different sensors are sent to the central database in a predefined amount of time. Another author proposed a parking system with intelligence that is created on the visual processing, the technique decides the accessibility of the space using deep learning. The proposed model is equated using the methods that are exit in the PKLot and CNRPark_EXT[10,11,12]. Other authors implemented a prototype that uses, RFID, sensor, and IoT to identify vehicle particulars and also uses IR sensor to locate the vehicle's existence, allowing all details to be accessible via IoT devices[13]. Pawowicz et al. [14] used an RFID-based technique to better handle traffic control in an urban areas, although the issue of predicting continues with this technique.

The authors of Giufr et al. [15] suggested an Intelligent Parking Assistant. Their intelligent parking management system is based mostly on the utilization of sensor networks.

However, this research ignores the technique of machine learning and the benefits of the Internet of Things. Another author uses Haar-AdaBoosting and CNN algorithms, Xiang et al. [16] suggested a method for the identification of real-time parking utilization at gas stations. Another deep learning case study [17] developed an automatic valet parking that is based on robotic valets which uses hybridization in smart parking that helps to maximize the use of parking area that uses the well-known technique that is Deep Q-Learning, a reinforcement learning. The authors of Camero et al.[18] (2018) introduced a new technique for processing parking occupancy rate predictions that uses deep learning by combining recurrent neural networks. The paper [19] provided a framework for designing an efficient parking control system that uses an innovative video processing technology that will help to assign vehicles to the vacant open parking slot at the entry point. This method is used for real-time forecasting of events. Stolfi et al. [20] use a study using parking occupancy data to examine a variety of forecasting methodologies that includes time series, Fourier series, k-means grouping, and polynomial alteration. The developed model is still critical and does not include rising technologies such as IoT and can be improved. Bachani et al.[21,32] provided a thorough examination of one of the most important components to develop an intelligent parking system, which includes the selection of appropriate sensors and optimal placement for precise detection.

In this paper, we propose an IoT-enabled smart parking model that will help in predicting the available parking space that fuses two technologies that is data mining and IoT technology.

III. METHODOLOGY

In the literature, many works are presented as smart parking system but fails to meet the major challenges. The issues which are to be at the limelight are user privacy, selection of parking slots in quick time, presenting an efficient and real-time considerable system, etc. In the current work major of the issues presented are considered for a solution. The current work presents a predictive system, which uses data mining techniques for fast and easy retrieval. For complete prediction proves Ensemble technique is considered. The entire system is designed considering two different types of agents as detection agent and the smart parking agent, which reacts as per the learning and retrieval results.

A. System Model of IoT Enabled Smart Parking System

There is the number of models available but the shortcoming of the available models is that a large amount of data is generated by the sensors to handle that data there is no specific model available that will handle a large amount of data. That is the reason we proposed multi-scanning model because sensors generate a large amount of data every second and we need to accumulate only that data which is important for parking-related information. The main idea behind making the smart parking system is to utilize the available parking space efficiently.

“The block diagram of the multi-scanning proposed methodology is shown in Fig. 1”.

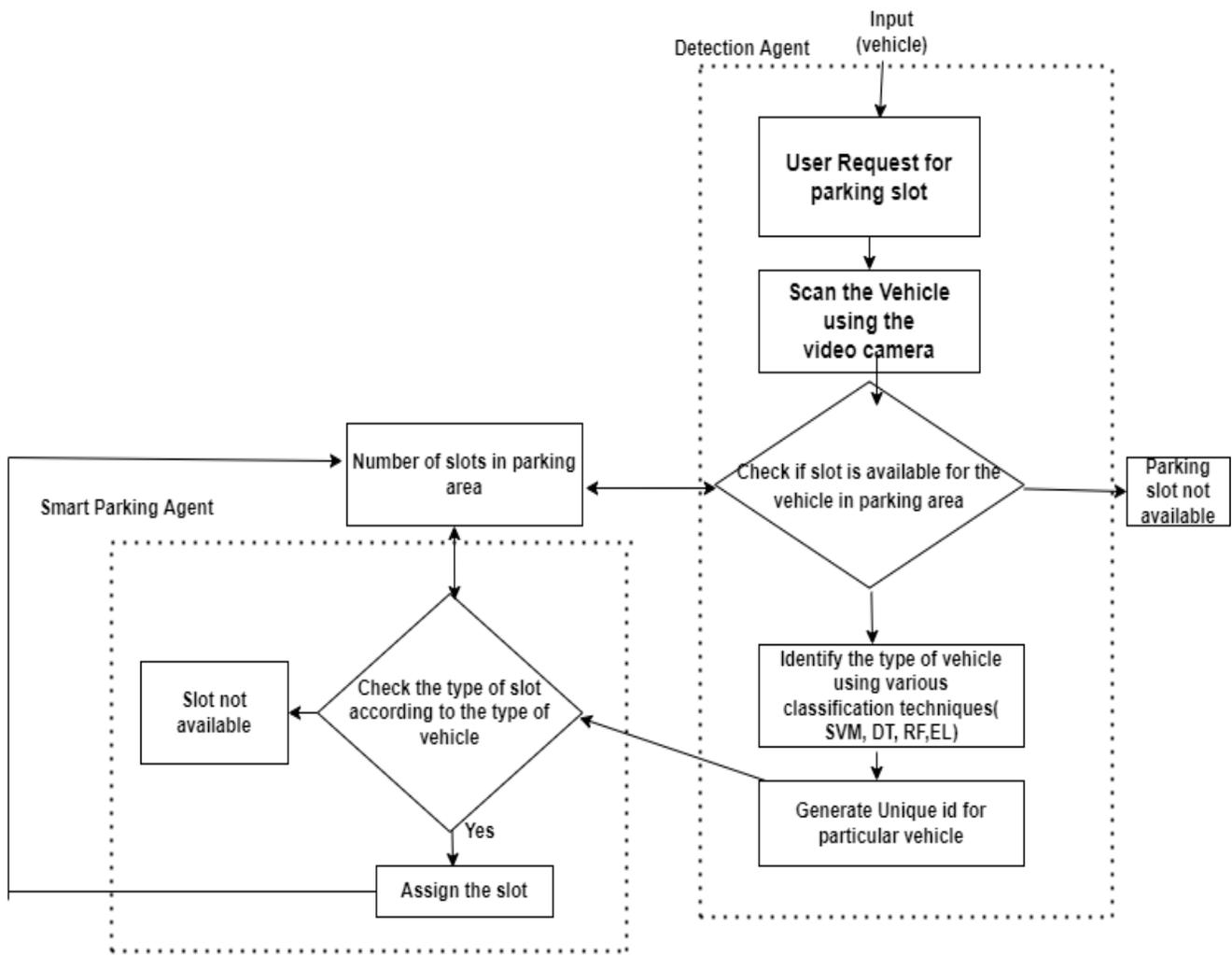


Fig. 1. Multi Scanning Model for IoT Enabled Smart.

The methodology is divided into two different parts:

- Detection agent.
- Smart parking agent.

The detection agent performs the identification of the type of vehicle. First, the user gives a request for an available parking slot. After getting the user's request, the system starts to detect the type of vehicle the sedan, coupe, SUV, bus, auto, etc., and assign a particular vehicle id to the specific vehicle. In the next part, the smart parking agent gives a particular location to the vehicle if the slot is available in the parking space. The slots of the parking areas are classified in advance by the smart parking agent according to the type of vehicle. Based on classified slots the smart parking agent will assign the available parking slot to the vehicle. The classifier that will help the smart parking agent is the ensemble technique.

The IoT-enabled smart parking model is a smart structure that uses a detection method to determine the parking space availability to assist the drivers with the accessibility of the parking space. A smart parking system needs the following elements:

- Sensors.
- Smart mobile application.
- Database.
- Gateway Hardware.

B. Ensemble Predictive Technique

This classifier combines several basic techniques such as decision tree or random forest. Every basic model provides an alternative approach to predicting the datasets and the final forecasting is the result of the combination of the basic models that will enhance the accuracy of the results. Because the uncertainty is significantly smaller than that provided by one of the individual base models that make up the overall model, the prediction technique of combining the predictions of a collection of individual base models often provides for more consistent and accurate output prediction. Therefore, the final ensemble-based model rectifies the individual errors generated by the fundamental models, resulting in a significant reduction in total error. The basic models should be required to meet two characteristics to be effective: they must be independent and weak models.

The original plan was to split the training data D into n basic data sets to train n models (m1;m2,..., mn). When n becomes high, however, this strategy is eventually outperformed because it promotes under-fitting. To deal with this kind of limitation various approaches are used that will re-sample the training data into n independent and greater data sub-samples to develop weak models. The number of techniques has been employed to accomplish it, the most well-known of which include bagging and boosting. As a result, the suitable algorithms for such kind of situations are volatile algorithms that include decision trees and neural networks that are modified during data set to produce a different model. A flow chart of a prediction system based on a set-based model is shown in Fig. 2.

The next sections covers our basic model which is decision trees and also contain two sampling strategies bagging and boosting and three-set models that contain the Random Forest regressor method (RFRM), Gradient boosting method (GBM), and Adaboost regressors(ARM) methods.

C. Decision Tree

The decision tree is used to depict the repeated division of the original entire space in the prediction tree technique. The terminal node or the leaves nodes represents a partition cell and is linked with a basic model that will apply only to that particular cell. For better understanding, think about a regress equation with two independent variables X1 and X2, and a continuous output variable Y. The concept is to divide the space into two sections and model the Y response (mean of Y) in every area separately. We again divide each section into two more sections and repeat the procedure till reaches a halt rule [22]. Each region's answer is frequently treated as a Cm constant.

$$\text{Minimize(SSE)} = \sum_{Ek=1}^{Em} \sum_{i=1}^n (yi - ck)^2 \tag{1}$$

$$\text{Minimize SSE} + \alpha|T| \tag{2}$$

The best Cm is simply the average of Yi in the Rm region because the optimization objective is to minimize the sum of squares [23]. Applied a cost complexity parameter (α) to identify the optimal region, which disciplines independent function in Equation (1) for leaf nodes of the tree T as given in Equation (2). The greedy algorithm is used to find the best division variable and split point. The optimal partition point for each splitting variable can be found by scanning all important parameters; it should be done fast. It is feasible to determine the finest pair of division parameters and division points by examining all of the input variables.

D. KNN for Regression

KNN uses non-parametric methods to resolve any regression problem, it is also known as lazy learner [24].If the incident values are categorical that it will assign the regression value by aggregating the nearest k neighbor values for quantitative occurrences or implementing the majority vote to the k neighbors. By aggregating the k neighbor outputs (constant weights) the values can be forecasted [25].

E. Ensemble Technique for Regression

The summary of the general idea is shown in Fig. 3 it shows that the model is based on three parts that is bootstrapping, intermediate model, and aggregation. Bootstrap split all the data D into n data D1 ,D2...Dn. Now we create an intermediate regressor Rj for each data set Dj and an aggregate of the successive regressors Ri will be the final regressor. The bootstrapping used in the Random Forest algorithm and the Boosting used by Gradient Boosting and Adaptive Boosting are two of the most powerful approaches derived from this fundamental principle.

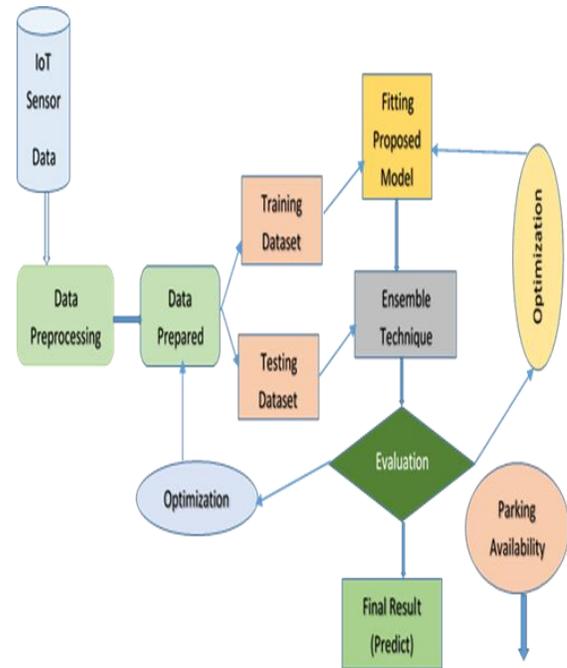


Fig. 2. Ensemble Technique for Prediction of Real Time Parking Space.

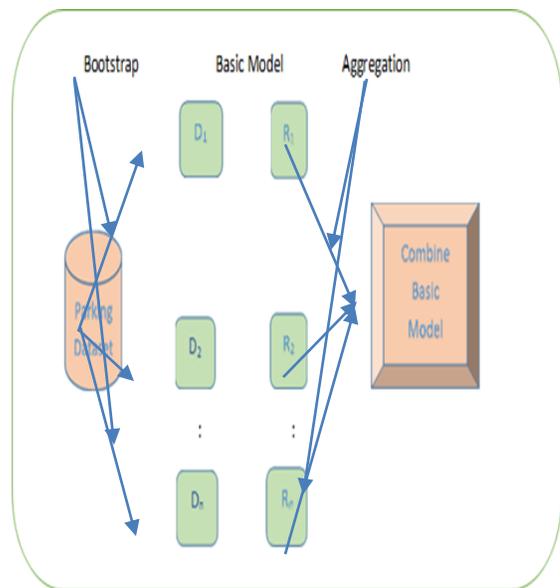


Fig. 3. Basic Working Approach of Ensemble Technique.

F. Bootstrap Aggregation / Bagging Technique

1) *Bagging Regression Method (BRM)*: Bagging term is derived from "Bootstrap Aggregating." The algorithms are shown in the context of regression and they can be easily extended as a supervised classes. Random vector describes the learning data is denoted by the values that is (X, Y) , where X is in R^q and Y is in R . $D_n = (X_1; Y_1); (X_n; Y_n)$ is a sample that is independent and evenly distributed. The regression function (X, Y) and $m(x) = E[Y|X=x]$ are both defined by the same law. The mean square error of an estimated m and its bias-variance representation for $x \in R^p$ is as follows:

$$(\hat{m}(x) - m(x))^2 = (E\hat{m}(x) - m(x))^2 + V(\hat{m}(x)) \quad (3)$$

These comprise of integrating a number Z of models $m_1; m_2, \dots, m_z$ in such a way that:

$$\hat{m}(x) = \frac{1}{Z} \sum_{k=1}^Z (m_k(x)) \quad (4)$$

And

$$E(m(x)) = E m_1(x) \quad (5)$$

And

$$V(m(x)) = 1/Z V(m_1(x)) \quad (6)$$

As a result, the composite model's bias is the same as the m_k , but the variable reduces. By generating aggregated models on bootstrap samples, the bagging method attempts to reduce the correlation between them.

2) *Random Forest Regression Method (RFRM)*: It is a specific bagging approach that involves a collection of trees that is based on random factors. The classification and regression tree algorithm, whose concept is to iteratively split the space formed by the independent variable reciprocally, is most commonly used to create trees. For each phase of the partitioning, a section of the space is divided into two sub-parts based on a variable X_j [23].

G. Boosting Method

The basic idea behind this method is to create a group of models that are combined using a weighted average of predictions. In contrast to the previous bootstrap aggregation, the standard increase in subgroup formation is not random. The main impression behind boosting algorithms is to train forecasters in order, with everyone attempting to accurate the one before it [26,27]. Since every new subdivision is repeated on the former comprised features that could have been mislabeled by prior models, so the performance depends on the performance of preceding models. Most general boosting methods are AdaBoost and Gradient Boosting.

1) *AdaBoost Method (ABM)*: It is also known as the adaptive boosting method. It is centered on the assumption that a new forecaster pays a bit more recognition to the training sample under which the antecedent has adjusted to rectify the fault of its precedent. The outcome is the new predictor whose main focus is more on difficult cases. Let's take an example, Assume a predictor which is nothing more

than a decision tree for creating an AdaBoost classifier. Predictions on the set of patterns are made using a fundamental tree structure. After that, the weight assigned to the misclassified training instances is increased. The altered weights are then used to create a second classifier [28]. The second classifier predicts the outcome of the training game once more. After that, the weights are adjusted and the process will continue until all the predictors are in order, the set makes predictions in the same way as bagging do. The primary difference is that the weights of the generated predictors are dependent on their total accuracy calculated over the weighted training set [29].

2) *Gradient Boosting Method (GBM)*: Gradient Boosting is similar to AdaBoost in that it gradually adds predictors to a set, with each one attempting to rectify the faults of the one before it. Instead of modifying the class weights at each repetition as AdaBoost does, this method seeks to match the new classifier to the prior one's residual errors [29].

IV. IMPLEMENTATION RESULTS

The given section, analyzes and discusses the performance of the proposed methodology using a data mining classifier that is Ensemble technique. Section A, discuss about the dataset.

A. Dataset

Fig. 4 shows the real-time prediction of parking space using the ensemble technique. Sensors are installed in different parking areas in smart cities. These sensors are connected to the database that collects the IoT data with the help of sensors installed in different places. In this paper, data was collected using Birmingham in (U.K) dataset as a CSV file format. Parking sensors [30] collect over 6 months of data. The process of selecting relevant data involves removing inappropriate and duplicate information [30]. The Parking dataset consists following features.

- Code Number: It is described by an alphanumeric value that will identify the parked vehicle.
- Last Modification: It provides the time and date of the most recent updation for every parking block's availability data. Between 9:00 a.m. and 5:00 p.m., schedules are logged.
- Size: It contains the capability of each parking space.
- Occupancy: It includes every vehicle park's activities that are modified every 30 minutes.
- Status: It represents the status of the parking space i.e. free or available.

We created a particular feature called the accessibility rate (AR) from these characteristics, which is the ratio of the availability excluding the occupancy on the parking areas space at time t of the dated d . In our scenario, we use the following formula to calculate it:

$$AR_q(t) = \frac{Capacity_q - Occupancy_q(t)}{Capacity_q} \quad (7)$$

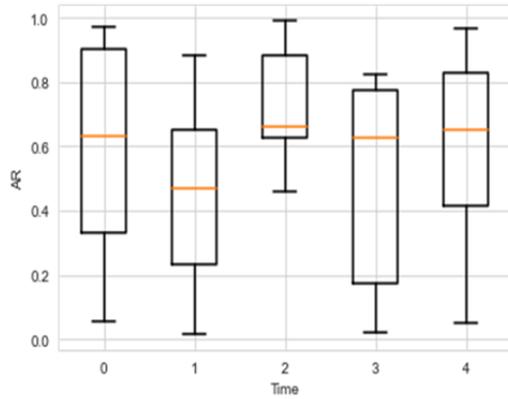


Fig. 4. Scattering of Availability of Parking Area as a Function of Time.

For reference, car park "BHMBCCMKT02" has a capacity of 455 parking spaces and an occupancy of 51 on 05/11/2020 at 06:49:32, thus we may infer our occupancy rate at this moment is 51/455, or 10.53 percent. The accessibility rate in the parking lot at the present time of day is 88.42%, indicating that the user can park here without causing traffic congestion. Fig. 4 illustrates the rate of parking availability as a function of time. Having the availability rates in real-time we are now applying data mining algorithms to forecast the future availability rates without knowing occupancy.

B. Performance Measure

We evaluated the performance of the model using two primary parameters to develop an ideal approach: the r-square (R2) and the mean absolute error (MAE). In various aspects, the two terms can be used to assess the gap between actual and predicted parking availability rates. The following formula is used [31].

$$R^2 = 1 - \frac{\sum_{j=1}^N (AR_{j,p} - AR_j)^2}{\sum_{j=1}^N (AR_{j,p} - AR_j)^2} \quad (8)$$

$$MAE = \frac{\sum_{j=1}^N |AR_{j,p} - AR_j|}{N} \quad (9)$$

Here N stands for a complete number of cases, AR_{j,p}, it is the predicted accessibility rate of the case j & AR_j is the actual available rate in the particular event. The use of a single metric may not always allow the models to be distinguished. The absolute error may reflect the effect of precision in the prediction of the waiting time, and R2 would also demonstrate to us the percentage of the exact waiting time that has been successfully predicted, whereas the MAE will show the error characterized by the deviation and mean between the predicted and the real by choosing the effects of high variations. The consistency of these two measures will lead to the best model.

C. Optimization of Hyper Parameters

Here we will compare the performance of different mining techniques discussed in the section for the prediction of available space in the parking lot. We separated our dataset as discussed in the section into two sets: a training set (80%) for building the models and a testing set (20%) for evaluating the model and comparing their performance. To begin, we searched for factors that would allow us to construct reliable and efficient models. It is important to determine the number of

iterations, the learning rate, and the optimal weak learner for the boosting procedure. The number of repetitions, the out-of-bag error, and the best weak learner is the most critical factors in bagging algorithms.

The tests that were conducted to choose these variables using a basic decision tree as a weak learner are shown in the figure.

The appropriate parameters for the boosting method are shown in Fig. 5. We can see from Fig. 5 that, how our model works on optimized parameters. The default value learning rate is 0.1, and the effective rate for Gradient Boosting is about 0.4. For AdaBoost, the same frequency might be used. The appropriate number of predictions for Gradient Boosting is at least 200, although it is not as necessary for Adaboost (about 30 predictor variables). Beginning with the estimation, a persistent optimum classifier based on bagging is created. The out-of-bag error is nearly zero. We can observe from the graph that the distribution of BRM followed by RFRM is more suitable for greater performance than GBM and ABM, which are both less efficient.

1) *Result analysis of total dataset of parking area:* In this, we evaluate the result on the total dataset of all the parking areas. To evaluate overall parking capacity, we used the attribute "Code Number" that is used for giving the identity numbers ranging from 0 to 25 for the 24 parking slots. The evaluations involved assessing and making comparisons of four different approaches. In the bagging algorithm, we use an optimization approach that is (BRM and RFRM) and boosting algorithm uses (GBM and ABM) approach. Table I shows the results, from which we created in Fig. 6 and Fig 7, which depict the comparison results. In comparison to the boosting method, the bagging methods (BRM and RFRM) produce the best results for all two metrics.

MAE produces 0.000771 & BRM produces 0.999951 here it is clear that BRM gave an optimal performance. In respect of R2, both BRM and RFRM achieved a near-perfect score of 0.999951. These results are somewhat better than RFRM, with an improvement of less than 0.00006 for these two metrics.

MAE and R2 received expert evaluations of 0.024523, and 0.982871 respectively, from GBM. These results were superior to ABM's 0.087872 and 0.834494 for these metrics, respectively. The findings of the two boosting methods evaluated (GBM and ABM) were substantially different from those of the very similar bagging methods, as demonstrated by the trend in Fig. 6 and 7. For the tests conducted, boosting approaches were slightly less effective than bagging approaches.

TABLE I. COMPARISON OF PERFORMANCE

		MAE	R ²
Bagging	BRM	0.000771	0.999951
	RFRM	0.000783	0.999949
Boosting	ABM	0.087872	0.834494
	GBM	0.024523	0.982871

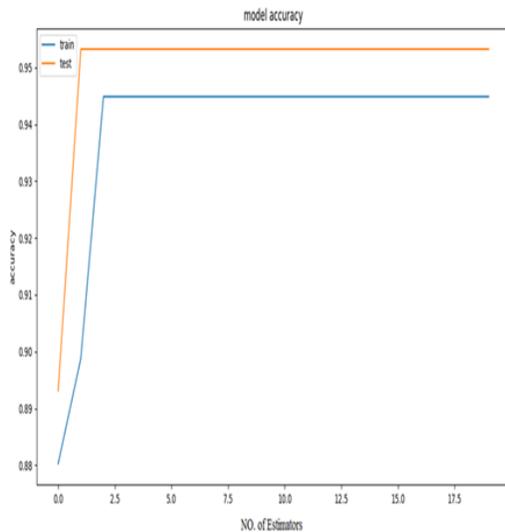


Fig. 5. Optimized Parameters of Boosted Method.

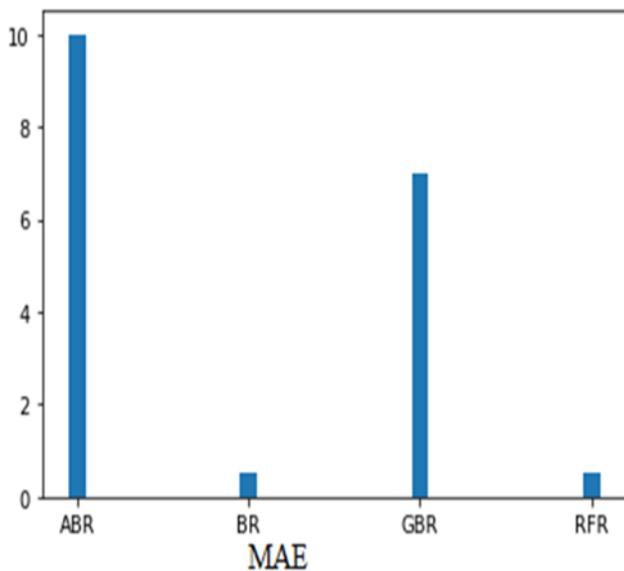


Fig. 6. Comparison of Performance using Mean Absolute Error (MAE).

D. Result Analysis

To validate the algorithms, we performed two tests: the first on the total dataset of all parking areas, and another on every parking area data.

1) *Result analysis of every parking area:* We put our approaches to the test to see if we could estimate parking availability in each area separately. Comparison to earlier

work that simply employed one loss function. In Table II we compare the results using two loss functions that are Mean absolute error and r- square (MAE, and R2). We also compare the minimum, maximum and average performance (i.e. R2, MAE) values in the same table.

2) *Comparison of MAE of predicted value with each parking space with previous work:* MAE values of the various approaches, the RFRM technique finds the smallest mean value of 0.00055. For R2 we observe that BRM gave the optimal mean value i.e. 0.99980 compared to other approaches. As a result, we reached the conclusion that optimizing using the bagging method, particularly BRM, was the most effective way of estimating the available space in each parking space.

The fundamental goal of prediction is to forecast values that are close to accurate as feasible, we compared the results of our method to those achieved in earlier studies (Table III). In the previous work, only one measure is used i.e. MAE. The comparison table depicts that our best method (BRM) reduced mean absolute error by 7.6% on average when compared to RNN [18] and by more than 6.8 percent when compared to [20], which used and compared to time series(TS), Fourier series (F), k-means clustering (KM), shift and phase modifications (SP), polynomial(P), polynomial fitted by k-mean centroid (KP). Furthermore, the standard deviation is also low using BRM (0.00023) when compared to earlier work, which had a minimum of at least 0.026. Even improved, when compared to earlier algorithms, BRM proves to be faster.

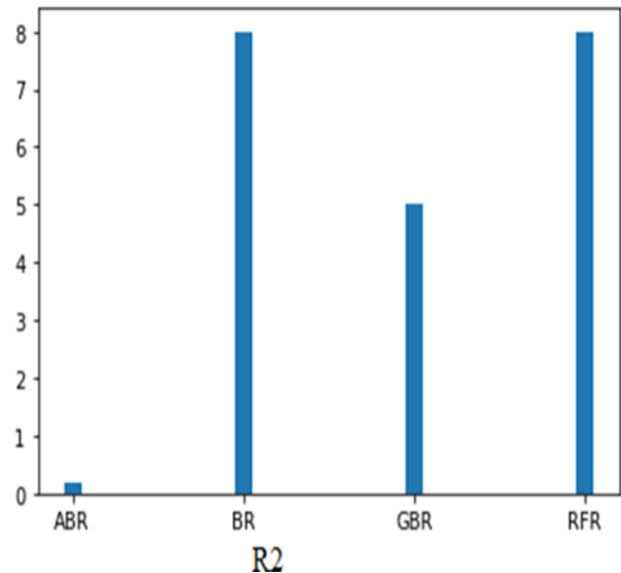


Fig. 7. Comparison of Performance using r-square (R²).

TABLE II. MAE AND R2 PREDICTION AVAILABILITY IN EACH PARKING

Vehicle ID	R ²				MAE			
	BRM	ABM	GBM	RFRM	BRM	ABM	GBM	RFRM
BHMBCCMKT02	0.9996	0.9951	0.9996	0.9996	0.0004	0.0090	0.0013	0.0005
BHMBCCMKT04	0.9996	0.9945	0.9996	0.9996	0.0005	0.0013	0.0012	0.0007
BHMBCCMKT05	0.9996	0.9972	0.9996	0.9996	0.0005	0.0092	0.0013	0.0005
BHMBCCMKT08	0.9996	0.9972	0.9996	0.9996	0.0004	0.0110	0.0011	0.0006
BHMBCCMKT10	0.9996	0.9971	0.9997	0.9996	0.0006	0.0142	0.0013	0.0004
BHMBCCMKT12	0.9997	0.9973	0.9996	0.9997	0.0006	0.0106	0.0015	0.0004
BHMBNCMKT15	0.9998	0.9962	0.9997	0.9996	0.0004	0.0060	0.0009	0.0007
BHMBCCMKT20	0.9972	0.9907	0.9962	0.9960	0.0012	0.0020	0.0012	0.0007
BHMBPKMKT23	0.9996	0.9961	0.9996	0.9996	0.0004	0.0145	0.0009	0.0007
BHMBCCMKT26	0.9997	0.9962	0.9997	0.9997	0.0003	0.0154	0.0012	0.0006
BHMBCCMKT01	0.9998	0.9900	0.9998	0.9996	0.0003	0.0150	0.0012	0.0005
BHMBCCMKT07	0.9989	0.9971	0.9985	0.9963	0.0006	0.0090	0.0014	0.0006
BHMBCCMKT09	0.9996	0.9956	0.9995	0.9965	0.0012	0.0080	0.0012	0.0005
BHMBCCMKT34	0.9971	0.9992	0.9970	0.9996	0.0004	0.0100	0.0009	0.0007
BHMBCCMKT27	0.9989	0.9941	0.9985	0.9981	0.0006	0.0061	0.0008	0.0006
NIA Park C26	0.9996	0.9951	0.9996	0.9996	0.0012	0.0101	0.0012	0.0005
NIA South	0.9996	0.9981	0.9996	0.9996	0.0005	0.0153	0.0012	0.0005
NIA North	0.9996	0.9955	0.9996	0.9996	0.0004	0.0170	0.0009	0.0005
BHMBDDMYT29	0.9996	0.9919	0.9994	0.9991	0.0012	0.0170	0.0010	0.0006
BHTBFFDTYK37	0.9996	0.9957	0.9994	0.9991	0.0005	0.0080	0.0005	0.0004
BHTBFFDTYK40	0.9996	0.9932	0.9996	0.9996	0.0004	0.0082	0.0008	0.0005
BHTBFFDTYK45	0.9996	0.9972	0.9996	0.9996	0.0003	0.0045	0.0012	0.0006
BHTBFFDTYK48	0.9996	0.9956	0.9996	0.9996	0.0005	0.0076	0.0009	0.0005
BHTBFFDTYK42	0.9996	0.9919	0.9996	0.9996	0.0004	0.0003	0.0058	0.0004
mean Value	0.99980	0.99471	0.99978	0.99972	0.00056	0.00976	0.00111	0.00055
max Value	0.99996	0.99819	0.99995	0.99996	0.00143	0.01812	0.00162	0.00152
min Value	0.99718	0.98563	0.99749	0.99642	0.00029	0.00208	0.00055	0.00028
standard deviation	0.00051	0.00298	0.00041	0.00050	0.00023	0.00428	0.00025	0.00024

TABLE III. COMPARISON WITH PREVIOUS WORK

Vehicle ID	SP	KPF	KM	F	P	TS	RNN	BRM	ABM	GBM	RFRM
BHMBCCMKT02	0.032	0.051	0.086	0.085	0.058	0.066	0.062	0.0004	0.0090	0.0013	0.0005
BHMBCCMKT04	0.067	0.071	0.147	0.148	0.082	0.112	0.136	0.0005	0.0013	0.0012	0.0007
BHMBCCMKT05	0.123	0.142	0.18	0.147	0.138	0.068	0.116	0.0005	0.0092	0.0013	0.0005
BHMBCCMKT08	0.123	0.143	0.136	0.132	0.122	0.087	0.102	0.0004	0.0110	0.0011	0.0006
BHMBCCMKT10	0.133	0.149	0.147	0.148	0.132	0.094	0.132	0.0006	0.0142	0.0013	0.0004
BHMBCCMKT12	0.078	0.115	0.123	0.121	0.096	0.087	0.110	0.0006	0.0106	0.0015	0.0004
BHMBNCMKT15	0.048	0.087	0.112	0.110	0.073	0.058	0.075	0.0004	0.0060	0.0009	0.0007
BHMBCCMKT20	0.055	0.056	0.086	0.084	0.035	0.041	0.076	0.0012	0.0020	0.0012	0.0007
BHMBPKMKT23	0.067	0.034	0.065	0.062	0.066	0.067	0.061	0.0004	0.0156	0.0009	0.0007
BHMBCCMKT26	0.076	0.084	0.073	0.071	0.083	0.128	0.085	0.0003	0.0154	0.0012	0.0006
BHMBCCMKT01	0.087	0.083	0.15	0.126	0.072	0.033	0.049	0.0003	0.0150	0.0012	0.0005
BHMBCCMKT07	0.086	0.057	0.086	0.083	0.035	0.071	0.071	0.0006	0.0090	0.0014	0.0006

BHMBCCMKT09	0.048	0.119	0.080	0.077	0.067	0.081	0.101	0.0012	0.0080	0.0012	0.0005
BHMBCCMKT34	0.039	0.079	0.14	0.153	0.119	0.073	0.123	0.0004	0.0101	0.0009	0.0007
BHMBCCMKT27	0.029	0.108	0.065	0.065	0.089	0.057	0.075	0.0006	0.0061	0.0008	0.0006
NIA Park C26	0.090	0.024	0.143	0.14	0.084	0.055	0.089	0.0012	0.0101	0.0012	0.0005
NIA South	0.033	0.050	0.173	0.071	0.041	0.034	0.100	0.0005	0.0153	0.0012	0.0005
NIA North	0.067	0.090	0.113	0.112	0.101	0.074	0.033	0.0004	0.0170	0.0009	0.0005
BHMBDDMYT29	0.032	0.055	0.048	0.049	0.028	0.054	0.079	0.0012	0.0170	0.0010	0.0006
BHTBFFDXYK37	0.092	0.089	0.102	0.101	0.073	0.067	0.053	0.0005	0.0080	0.0005	0.0004
BHTBFFDXYK40	0.057	0.075	0.064	0.064	0.031	0.078	0.091	0.0004	0.0082	0.0008	0.0005
BHTBFFDXYK45	0.083	0.119	0.119	0.121	0.05	0.095	0.049	0.0003	0.0045	0.0012	0.0006
BHTBFFDXYK48	0.016	0.084	0.081	0.92	0.089	0.061	0.033	0.0005	0.0076	0.0009	0.0005
BHTBFFDXYK42	0.035	0.054	0.065	0.066	0.032	0.032	0.037	0.0004	0.0003	0.0058	0.0004
Mean value	0.068	0.078	0.102	0.101	0.073	0.067	0.079	0.00056	0.00976	0.00111	0.00055
Max value	0.122	0.147	0.177	0.179	0.139	0.129	0.137	0.00143	0.01812	0.00162	0.00152
Min value	0.015	0.024	0.148	0.049	0.025	0.023	0.033	0.00029	0.00208	0.00055	0.00028
standard deviation	0.030	0.035	0.035	0.035	0.033	0.026	0.028	0.00023	0.00428	0.00025	0.00024

V. CONCLUSION

Among the important features of smart cities that directly help residents on day to day, basis is urban transportation. One of the common problems in urban areas is congestion which is intensified by the search for free parking spaces by at least 35%. The capability to predict the available space for parking the vehicle in the urban cities is a big challenge and the smart solution will significantly lower traffic jams and also reduces metropolitan pollution. Some writers presented methodologies and models for the prediction of available space for parking the vehicle have several limitations. The proposed model helps the users to predict the parking space in advance using smart devices. As we know lots of data is generated every second it filters the data according to the need of the user and provides valuable information in less time using limited memory space. IoT enabled smart parking predictive model should allow us to take advantage of all of the interconnected devices in smart parking lots for the collection of data, evaluate it, and communicate the results with the users. Data mining technique that is Ensemble predictive analytic, enhanced the prediction of free space available in smart parking areas.

REFERENCES

- [1] C. Stracener, Q. Samelson, J. MacKie, M. Ihaza, P. A. Laplante, and B. Amaba. 2019. The internet of things grows artificial intelligence and data sciences. *IT Profess.* 21, 3 (2019), 55–62. <https://doi.org/10.1109/MITP.2019.2912729>.
- [2] Z. Allam and Z. A. Dhunny. 2019. On big data, artificial intelligence and smart cities. *Cities* 89 (2019), 80–91.
- [3] M. Mohammadi and A. Al-Fuqaha. 2018. Enabling cognitive smart cities using big data and machine learning: Approaches and challenges. *IEEE Commun. Mag.* 56, 2 (2018), 94–101.
- [4] Trista Lin, Hervé Rivano, and Frédéric Le Mouél. 2017. A survey of smart parking solutions. *IEEE Trans. Intell. Transport. Syst.* 18, 12 (2017), 3229–3253.
- [5] Philippe Esling and Carlos Agon. 2012. Time-series data mining. *ACM Comput. Surv.* 45, 1 (2012), 12.
- [6] Liangbo Qi, Hui Yu, and Peiyan Chen. 2014. Selective ensemble-mean technique for tropical cyclone track forecast by using ensemble

prediction systems. *Quart. J. Roy. Meteorol. Soc.* 140, 680 (2014), 805–813.

- [7] Huai-zhi Wang, Gang-qiang Li, Gui-bin Wang, Jian-chun Peng, Hui Jiang, and Yi-tao Liu. 2017. Deep learning based ensemble approach for probabilistic wind power forecasting. *Appl. Energy* 188 (2017), 56–70.
- [8] Yoo, Seong-eun, Chong, Poh Kit, Kim, Taehong, et al., 2008. PGS: Parking Guidance System based on wireless sensor network. In: 2008 3rd International Symposium on Wireless Pervasive Computing. IEEE, pp. 218–222.
- [9] Kumar, Rakesh, Chilamkurti, Naveen K., Soh, Ben, 2007. A comparative study of different sensors for smart car park management. In: 2007 International Conference on Intelligent Pervasive Computing (IPC 2007). IEEE.
- [10] Almeida, De, Paulo, R.L., et al., 2015. PKLot-A robust dataset for parking lot classification. *Expert Syst. Appl.* 42 (11), 4937–4949.
- [11] Amato, Giuseppe et al., 2016. Car parking occupancy detection using smart camera networks and deep learning. In: 2016 IEEE Symposium on Computers and Communication (ISCC). IEEE.
- [12] Amato, Giuseppe et al., 2017. Deep learning for decentralized parking lot occupancy detection. *Expert Syst. Appl.* 72, 327–334.
- [13] Gandhi, B.K., Rao, M.K., 2016. A prototype for IoT based car parking management system for smart cities. *Indian J. Sci. Technol.* 9 (17), 1–6.
- [14] Pawowicz, B., Salach, M., Trybus, B., 2019. Infrastructure of RFID-based smart city traffic control system. In: Conference on Automation. Springer, Cham, pp. 186–198.
- [15] Giuffr, Tullio, Siniscalchi, Sabato Marco, Tesoriere, Giovanni, 2012. A novel architecture of parking management for smart cities. *Procedia-Soc. Behav. Sci.* 53, 16–28.
- [16] Xiang, Xuezhi et al., 2017. Real-time parking occupancy detection for gas stations based on Haar-AdaBoosting and CNN. *IEEE Sens. J.* 17 (19), 6360–6367.
- [17] Shoeibi, N., Shoeibi, N., 2019. Future of smart parking: automated valet parking using deep Q-Learning. In: International Symposium on Distributed Computing and Artificial Intelligence. Springer, Cham, pp. 177–182.
- [18] Camero, A., Toutouh, J., Stolfi, D.H., Alba, E., 2018. Evolutionary deep learning for car park occupancy prediction in smart cities. In: International Conference on Learning and Intelligent Optimization. Springer, Cham, pp. 386–401.
- [19] Mago, N., Kumar, S., 2018. A machine learning technique for detecting outdoor parking. *Int. J. Eng. Technol.* 7 (2.30), 39–43.
- [20] Stolfi, D.H., Alba, E., Yao, X., 2017. Predicting car park occupancy rates in smart cities. In: International Conference on Smart Cities. Springer, Cham, pp. 107–117.

- [21] Bachani, Mamta, Qureshi, Umair Mujtaba, Shaikh, Faisal Karim, 2016. Performance analysis of proximity and light sensors for smart parking. *Proc. Comput. Sci.* 83, 385–392.
- [22] Faris, H., Abukhurma, R., Almanaseer, W., Saadeh, M., Mora, A.M., Castillo, P.A., Aljarah, I., 2019. Improving financial bankruptcy prediction in a highly imbalanced class distribution using oversampling and ensemble learning: a case from the Spanish market. *Progr. Artif. Intell.*, 1–23.
- [23] Zhang, Y., Haghani, A., 2015. A gradient boosting method to improve travel time prediction. *Transp. Res. Part C: Emerging Technol.* 58, 308–324.
- [24] Anchal, P. Mittal, 2021. Evaluation of Data Mining Techniques and Its Fusion with IoT Enabled Smart Technologies for Effective Prediction of Available Parking Space. Vol 12, Number 4, 187-197.
- [25] Sinta, D., Wijayanto, H., Sartono, B., 2014. Ensemble k-nearest neighbor's method to predict rice price in Indonesia. *Appl. Math. Sci.* 8 (160), 7993–8005.
- [26] Freund, Y., Schapire, R.E., 1996. Experiments with a new boosting algorithm. In *icml* vol. 96, pp. 148–156.
- [27] Freund, Y., Schapire, R., Abe, N., 1999. A short introduction to boosting. *J.-Jpn. Soc. Artif. Intell.* 14 (771–780), 1612.
- [28] Mishra, S., Mishra, D., Santra, G.H., 2017. Adaptive boosting of weak regressors for forecasting of crop production considering climatic variability: an empirical assessment. *J. King Saud Univ.-Comput. Inf. Sci.*
- [29] Géron, A., 2017. *Hands-on Machine Learning with Scikit-Learn and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems.* O'Reilly Media Inc..
- [30] Camero, A., Toutouh, J., Stolfi, D.H., Alba, E., 2018. Evolutionary deep learning for car park occupancy prediction in smart cities. In: *International Conference on Learning and Intelligent Optimization.* Springer, Cham, pp. 386–401.
- [31] Xu, H., Ying, J., 2017. Bus arrival time prediction with real-time and historic data. *Cluster Comput.* 20 (4), 3099–3106.
- [32] Zheng, Yanxu, Rajasegarar, Sutharshan, Leckie, Christopher, 2015. Parking availability prediction for sensor-enabled car parks in smart cities. In: *2015 IEEE Tenth International Conference on Intelligent Sensors, Sensor Networks and Information Processing (ISSNIP).* IEEE.

An Optimized Kernel MSVM Machine Learning-based Model for Churn Analysis

Pankaj Hooda, Pooja Mittal

Department of Computer Science and Applications, Maharishi Dayanand University, Rohtak, India

Abstract—Customer churn is considered as a significant issue in any industry due to various services, clients, and commodities. A massive amount of data is being created from e-commerce services and tools. Analytical data and machine learning-based approaches have been implemented and utilized for CA (churn analysis) to design a plan, i.e., required to comprehend the rationale for the CC (Customer Churn) and to generate a profitable and actual customer holding program. The analytics and machine learning approaches mainly focus on customer profiling, CC classification, and detection of features that affect churn. However, there are no specific techniques which can be used to determine how often a prospective customer is inclined to cover all the expenses whether they are churned or not. In this paper, an Optimized Kernel MSVM classification model is proposed to predict and classify churn. In the proposed work, MSVM algorithm has been used for classification. The kernel PCA and ALO optimizer method has been used for Feature extraction and selection. The proposed model Optimized Kernel MSVM has been implemented on Tele-communication sector customer churn database to demonstrate the proposed model's generalization ability. The Optimized Kernel MSVM model has achieved an accuracy of 91.05%, AUC 85% being maximum and reduced the RMSE score to 2.838. The implementation shows that both churn detection and classification may be examined at the same time while maintaining the highest overall accuracy and AUC.

Keywords—CA (Churn Analysis); CC (Customer Churn); OKMSVM (Optimized Kernel-MultiClass Support Vector Machine) Model; KPCA (Kernel Principle Component Analysis); A.L.O. (Ant Lion Optimization) method

I. INTRODUCTION

Identifying causes for client loss, evaluating customer retention, and recovering clients have become essential ideas for several businesses. Organizations do many studies and efforts to prevent losing clients in support of acquiring new clients. Due to the increasing green technologies, growing users, and valuation solutions. The unregulated and rapid development of this area has resulted in increased failure due to deception and technological challenges. As a result, creating new methodological approaches has become a necessity. This issue has become a pioneer in numerous studies in India's communications industry, experiencing massive consumer shortages. These are some of the uses of information retrieval: churn analysis, which is broadly employed across all sectors. Organizations need to develop initiatives to enhance customer satisfaction and formulate plans for improved customer retention by assessing which customers are likely to transfer providers. This research aims

to figure out how a telecommunication business loses consumers.

In the same way that causes are examined, determining which kinds of clients are abandoned is also studied [1]. There are different ways to define churn. The most important two ways are contractual churn and non-contractual churn. When a person does not extend their contract after the termination period has passed, this is known as contractual churn. Such churn starts when the customer loses interest in the products and comes to a point where reintegration is no longer feasible [2]. It's most commonly found in churn issues that occur whenever users terminate their savings accounts or move their wireless provider from one provider. Non-contractual churn is the second condition. Users can generally leave service without specific timelines in a non-contractual scenario. The customer operations team first establishes a churning state, after which a customer who meets that criterion is identified as a churned customer. The user's behavioral modification time is used to accomplish it. The individual is considered a churn client whenever the duration of idleness or changing behavior surpasses the limit.

The period is the interval established as the boundary of the inactive period throughout this operation [3]. There usually are two stages of churn prediction: (i) selection of features to evaluate the subsections of the characteristics that distinguish users would be preferable to locate the churned or not, and (ii) churn forecast. Customer churn is a helpful way to track how many clients are lost. Telecommunications firms frequently lose valued consumers and, as a result, revenue to competitors. The telecommunications business has seen significant transformations in recent decades, including the growth of different products, scientific innovations, and more competitiveness. Therefore, customer churn forecasting throughout the telecommunication sector has grown critical of players in the industry necessary to defend existing customer loyalty, maintain their competitive advantage, and enhance relationships with the customers [5]. Among the most complex difficulties in the telecommunication sector is supporting consumers with an elevated unemployment probability. Consumers usually choose churn choices due to the increased number of telecommunications companies and competitive pressures. As a result, telecommunications companies have realized the value of customer retention rather than obtaining new customers [16]. A variety of reasons influences customer churn. Prepaid clients, unlike post-paid subscribers, really aren't constrained by contractual arrangements; therefore, they frequently churn for the most little causes. As a result, predicting the customer churn rate is

challenging [27]. The additional aspect is client loyalty, which is influenced by the service providers' service and product performance. Customers may switch to a competitor with a more extensive range and higher transmission qualities due to broadband service and transmission reliability problems [17] [18]. Inadequate or unsatisfactory response to concerns and invoicing issues are other variables that increase clients' likelihood of emigrating to the opposition. Clients may transfer to the competitors based on shipping expenses, insufficient functionality, and obsolete equipment. Customers frequently evaluate suppliers and switch to whichever they believe offers a significantly better price. Even if it involves obtaining no potential clients, a telecommunications business can do OK, providing better care to established customers. The typical churn rate amongst telecommunication companies in the telecommunications business is around two per cent globally, resulting in a massive yearly loss of around \$1 trillion.

The research motivation to acquire the issue mentioned above, the corporation must accurately forecast the customers' behavior. There are two strategies for managing the churn rate: There are two types of reactions: (a) reactive and (b) proactive. In the reactive strategy, the corporation prepares for a reschedule from the user, following which it provides the appealing client decisions to maintain them. The likelihood of churning is predictable under the preventative approach, and clients present alternatives appropriately. It's a supervised learning [15] model in which churners and non-churners are differentiated. Machine learning, which comprises regression analysis, vector support network, RF, NB, LR, and others, has proven to be an extremely effective method for predicting data on the performance of formerly gathered information to address this challenge. Pre-processing and feature selection play an essential part in improving classification accuracy in machine learning techniques. Researchers have devised a slew of feature selection methods that can help minimize dimensionality, computational time, and overfitting. The supplied input sequence identifies the features relevant for predicting churn [4]. This research aims to group different consumers and identify the elements contributing to every category's turnover.

Furthermore, this study aims to design a churn prediction model using the OKMSVM model and apply it to predict churning consumers. This study will evaluate the information presented to establish that specific variables are associated with churning prediction. The number of variables gathered will be examined, and the system will be improved for the finished product. This study also seeks to determine churn prediction expense by determining the overall cost of clients who have churned to date and how much revenue can be avoided unless we can enhance our customer defection monitoring. Following the customer churn, a group will determine how maintenance strategies will be provided.

The organization of this work is as follows: The existing techniques of churn analysis are measured in section 2. In section 3, the existing issue and the problem is given. Sects 4 and 5 define the proposed work with a flow chart and result in an analysis with a detailed explanation. Sect 6 shows the conclusion and further work in the churn analysis.

II. RELATED WORK

In this section, existing methods for churn prediction are analyzed, and comparison tables are shown for better analysis. Jain et al. (2021) [6] proposed a multi-attribute strategic planning system combined with machine learning techniques. The name of the suggested strategy was the Worker's churn forecasting and retaining approach. A two-stage methodology was used for categorizing employees by creating an incredible achievement employment significance paradigm. The first proposal of the suggested methodology was to enhance the implementation of the entropy-based approach to allocating weighting factors to personnel achievements. Furthermore, for assessing the value of the personnel accomplishing and their class-based classification, an enhanced methodology (CatBoost) was implemented. The CatBoost method was then used to forecast employee churn by classification. Ultimately, based on the forecast findings and attribute rating, the authors had presented a retention strategy. Sarac, F., et al. (2021) [7] designed a two-level churn approach to evaluate whether a client would churn and determine how the consumer would pay for services. A classification technique called support vector machine was employed for such categorization component, and a recurring monthly cost was forecasted using machine learning-based support vector regression methodology. An autonomous feature selection technique, the multi-cluster attribute selection approach, was used to pick the most relevant features, including both evaluations. For uniformity, the same attribute selection strategy was employed in both evaluations to determine its effectiveness. The proposed scheme technique was then tested mostly on IBM. Telco Customer Churn set of data, which included over 7000 clients, to validate its relevance and generalization capabilities. Lalwani, et al., (2021) [8] proposed a machine learning-based approach. There are six steps to the proposed approach. Data pre-processing was the first step, and feature analysis was carried out during the second step. The third step used gravitation methodology to evaluate essential feature selection. The input was separated into two sections: training data and testing data, with an 80/20 proportion. On the training data, one of the most common estimation methods, such as LR (Logistic Regression, SVM (Support Vector Machine), Decision Tree (DT), etc., were implemented. The boosters, as well as ensemble approaches, were used for efficient predictive performance. Furthermore, K-fold cross-validation was performed over the training data for hyper parameter to minimize modeling fitting problems. Lastly, the confusion matrix and AUC curve were used to examine the test dataset outcomes. The Adaboost classification model was reported with 81.71 percent accuracy. Bayrak, A. T., et al., (2020) [9] proposed a churn estimation model with the help of an advanced learning technique. The designed model was based on LSTM (long short-term memory) technique. Clients' information was organized in a particular sequence in the customer information architecture. A long short-term memory design was produced using sequencing information to determine users' churn phases and therefore was compared with the existing categorization approaches. Including the assumptions, the suggested model achieved success and differentiated from the related research. Jain et al., (2020) [10] designed a framework for determining consumer attrition; the

proposed methodology was used two machine-learning methodologies: logistic boost and logistic regression. The testing was performed using the WEKA ML (machine learning) technology and an actual dataset from the Orange firm in the United States. Various assessment methods were used to display the results. Ullah, et al., (2019) [11] proposed study revealed churn characteristics, which were crucial in deciding churn's core origins. CRM might promote efficiency, offer suitable offers to talented churn clients associated with particular behavioral patterns, and drastically improve the corporation's advertising campaigns by identifying the main churn drivers from user information. The Receiving operating characteristic area, recall, Accuracy, Precision, and f-measure of the suggested churn estimation method were all examined. The findings demonstrate that by employing the R.F. method and k-means cluster formation, the suggested churn proposed method had obtained significant churn categorization and customer preferences. Alboukaey, et al., (2020) [19] suggested a regular churn forecasting-based model rather than quarterly churn forecasting, dependent on the user's regularly dynamic characteristics rather than his quarterly behavior. The authors expressed the everyday behavior of customers as multidimensional data and suggested four predictions based on the description to forecast the everyday turnover of the customers. A deep learning framework was suggested by Seymen, et al., (2020) [20] to determine if commercial consumers would churn in the later. The framework was validated against regression analysis and convolutional neural network approaches, both of which were usually applied in churn estimation analyses. Recall, A.U.C., and Precision evaluated the algorithms' outcomes with reliability classifiers. The analysis indicates that the trained model outperformed the other approaches in terms of prediction and classification. Hu, X., et al., (2020) [21] designed a machine learning-based

framework based on an integrated approach. The machine learning-based neural network and decision tree were used in the integrated approach. This work develops a composite churn prediction statistical model and tests its performance using statistical results. Ahmed, et al., (2019) [22] proposed a telecom churn prediction approach that integrates ensembles layering and uplifting-based techniques. Traditional performances and expense (cost) heuristics were used in the assessments, and expense heuristics received the most attention. The proposed methodology, operations had a high level of connection across performance metrics and business objectives, making the method suited for the majority of cost-sensitive operations. Yu, et al., (2018) [23] designed a methodology based on the machine learning technique and a particle categorization performance tuning back propagation (B.P.) networking for telephony customer churn estimation was proposed that periodically performs P.F.C. (Particle Fitness Computation) and PCO (Particle Classification Optimization). Abou el Kassem, et al., (2020) [24] and Khan, Y., Shafiq, et al., (2019) [25] developed estimation techniques for churn prediction with the help of machine learning approaches. Butgereit, et al., (2020) [26] used the machine learning technique to anticipate when consumers were poised to churn and while churning was examined. Such estimations were then applied to search through unstructured or semi-structured user input log files for explanations of why and how the user could be churning. In Table I, various existing methods for churn prediction are compared. The comparison is based on implementing methods, comparison techniques, and existing issues.

The existing methods of churn analysis with proposed parameters, comparison parameters are depicted in Table II. The conclusion of the current churn analysis with future enhancement is also described in Table II.

TABLE I. COMPARISON OF EXISTING METHODS OF CHURN PREDICTION

Author Name	Techniques	Proposed Method	Comparative Methods	Problems
Jain et al., (2021) [6]	Machine learning-based decision-making technique	Multiple attributes based decision-making Approach	LR (Logistic regression) DECISION TREE (DT) RANDOM FOREST (RF) CatBoost SVM (SUPPORT VECTOR MACHINE) XGBoost	More performance metrics are required for the reliability of the model
Sarac, F., et al., (2021) [7]	Two-level based hybrid support vector machine	Churn prediction based on machine learning methodology	SUPPORT VECTOR REGRESSION (SVR) SUPPORT VECTOR MACHINE (SVM)	More data will be collected for the reliability of the proposed model and need to reduce generalization issues
Lalwani et al., (2021) [8]	XGboost and Adaboost based classification model	Machine learning-based employee churn estimation approach	XGBoost LogGISTIC REGRESSION NAIVE BAYES CatBoost SVM (SUPPORT VECTOR MACHINE) RANDOM FOREST	Imbalanced datasets issues
Bayrak, A. T., et al., (2020) [9]	Long Short Term Memory-Based Model	Sequential statistics based LSTM (Long short term memory) technique	-	More Inaccurate predictions
Jain et al., (2020) [10]	Logit boost and Logistic regression-based model	Supervised learning based churn estimation technique for telecommunication	Fuzzy C-means technique Multilayer Perceptron Firefly based hybrid approach Support vector machine	Limited dataset for the reliability of the model
Ullah et al., (2019) [11]	Machine learning-based Random forest model	Random forest-based churn estimation in the telecommunication field	Multilayer perceptron (MLP) Random tree Logistic regression Naive Bayes	Insufficient results of churn prediction

TABLE II. EXISTING METHODS OF CHURN ANALYSIS: PARAMETERS, DATASET, CONCLUSION, FUTURE ENHANCEMENT

Author Name	Proposed Parameters	Datasets	Conclusion	Future Enhancement
Jain et al., (2021) [6]	Recall Accuracy Precision MCC (Matthew's correlation coefficient)	HRIS employee database	Using a multiple attribute judgment method, a unique incredible achievement personnel significance framework has been suggested to associates into several classifications based on the evaluation criterion	The proposed methodology will be implemented as a real-time approach
Sarac, F., et al., (2021) [7]	Root mean square error (RMSE) Accuracy	IBM telecommunication based dataset	The proposed methodology classifies the employees efficiently according to services charges	A deep learning-based approach will be implemented for more efficient outcomes
Lalwani et al., (2021) [8]	AUC Accuracy Recall F1-measure Precision	-	The gravitational search technique had provided effective results on feature selection and minimized the dataset dimensionality.	Based on deep learning and reinforcement learning proposed methodology will be improved in future
Bayrak, A. T., et al., (2020) [9]	Recall Precision F1-score	-	The proposed framework gives promising findings as well as efficient from other existing methods	The hybrid methodology will be implemented for more efficient results
Jain et al., (2020) [10]	MAE (MEAN ABSOLUTE ERROR) Kappa Statistic RMSE (Root Mean Square Error), Root relative square error Relative absolute error, Recall, Precision F-measure AUC	American Company Orange-based dataset	The proposed model employed two self-contained strategies that worked admirably; however, stand-alone procedures cannot incorporate all of the variables, improving the results.	A fuzzy theory-based framework will be implemented for effective outcomes
Ullah et al., (2019) [11]	Precision F-measure Accuracy Recall ROC	Telecommunication based dataset	For telecommunications company decision-makers, the proposed framework established client retention recommendations.	Lazy learning and eager learning approaches will be implemented in the future for efficient results

III. PROBLEM STATEMENT

Conventional customer churn estimation is based on corporate administrators' perspective that is being used inductive approach, so executives may anticipate turnover for existing clients relies on churned users' attributes.

Even so, expertise may be uncertain; specifically in the context of a problematic issue, no helpful instruction can also be provided solely based on expertise; because of a company's limited resources, finances should be decided to invest first in recapturing those clients with the most acceptable churn probability.

The traditional system of predicting that customers are willing to churn and which clients are much less inclined to turnover is ineffective.

As a consequence, if a company wants to make a logical forecast of churn prediction, it must employ numerical algorithms as well as "machines" to detect the connection among statistical features as well as customer churn, determine if clients are being churned, and will provide the churn probability [12].

In recent years, churn prediction has been a significant concern in the telecommunication sector. Telecom carriers must identify such clients before their churn to address this issue. As a result, creating a different classification that

accurately predicts churn is critical. This classification should determine clients who are likely to churn in the coming years, allowing the users to respond quickly with relevant deals and discounts.

Machine learning techniques for categorization, such as k-nearest neighbour, decision trees, logistical regression, neural networks, Nave Bayes, and others, are the most popular approaches for this objective. In addition, studies should concentrate on uncovering innovative capabilities that are the most successful in forecasting client attrition [13]. Various costs can cause customer churn, individual characteristics, information and service, facilitating conditions, economic indicators, promotional strategies, and competition' market participation. The best approach to recovering churned clients and lowering the churn rate is determining churn prediction's cause(s).

According to a brief overview of influential factors of churn prediction in recent times, academics' investigation on attempting to influence determinants of consumer churn in the telecommunications industry consists of three main components: 1st, usage factors like call period and usage portion, accompanied by quantitative client variables like personally identifiable information as well as maturity level, client revenue, and satisfaction of customers, as well as eventually, corporation relevant factors.

IV. RESEARCH METHODOLOGY

In this paper, a new model is proposed to analyze the churn. Fig. 1 shows various modules of the proposed model to perform churn analysis.

The initial input has different data elements collected from the dataset and considered labelled experiences for the machine learning model.

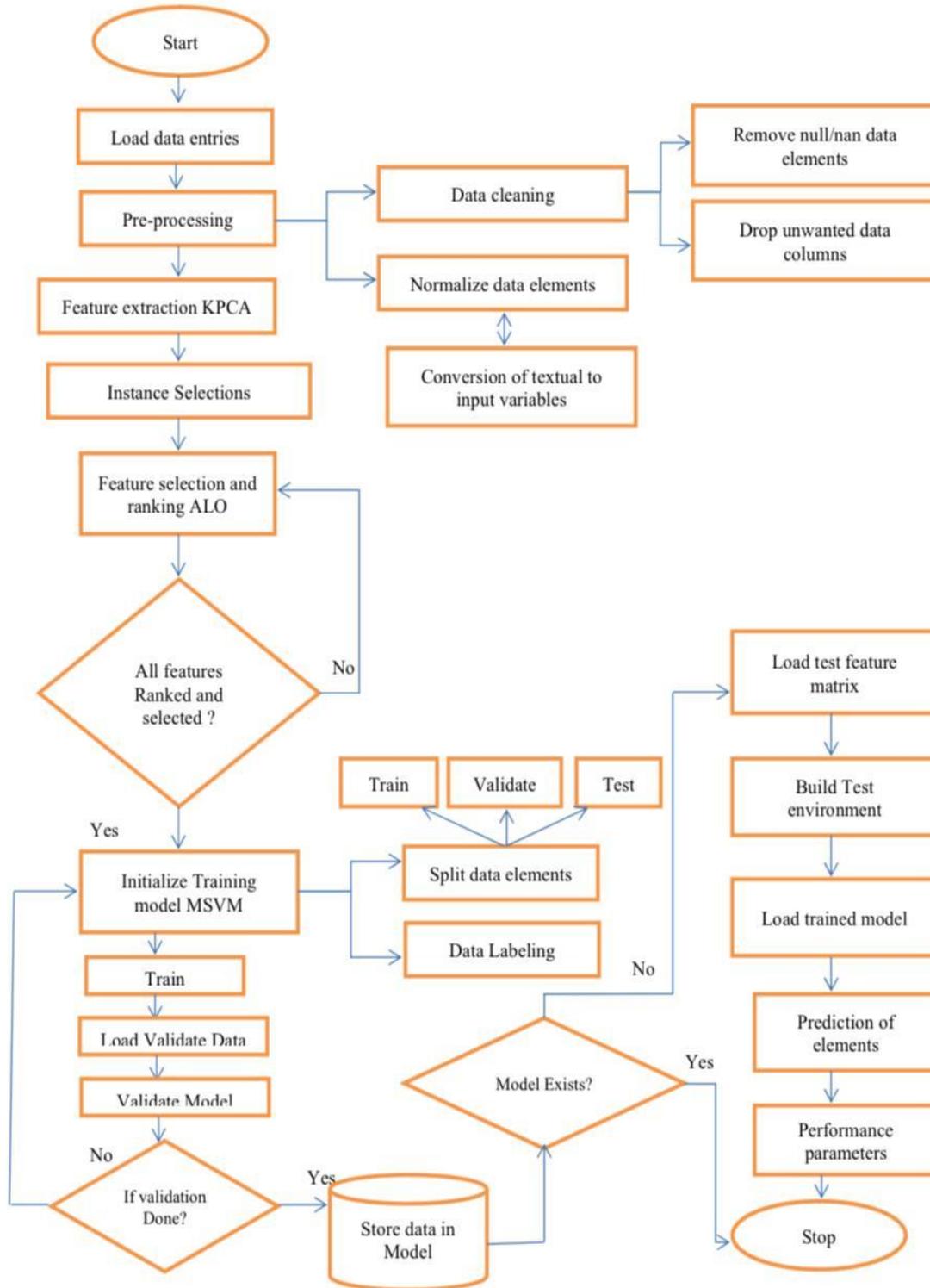


Fig. 1. Flowchart of Proposed Model OKMSVM.

These experiences are labelled data sources and help to create unique patterns for Churn analysis. The modules present in proposed model process the dataset and give the ability to model for further predictions.

The first phase of the proposed model is to pre-process the uploaded dataset. This process helps to clean the input data, remove all unwanted elements, remove or replace the String elements, remove non-processing data entries, etc.

The pre-processed data extracted the features and built unique feature patterns for training categories. The feature patterns combine numbers in the M*N matrix that form a unique combination against a particular case.

The feature extraction process is handled by the KPCA algorithm in the proposed architecture. KPCA algorithm is used to extract the features in a matrix. It reduces the dimensionality of data without much loss of data. It applied to the dataset that is linearly separable as compared to other methods.

KPCA utilizes a KF (kernel function) to project the database into a high dimensional feature space, linearly separable. Extracted features are used to process with the Ant Lion optimizer. ALO is an optimization module of the proposed architecture that helps reduce the feature set's error probability. The more miniature error probability training set provides more accuracy in the trained model. It is an iterative process to find the best cost solutions for input feature patterns.

Once all the cases are processed, the data element is processed with the initialization module of the prediction model. This module processes the optimized data elements and builds various subsets of the actual datasets. These subsets are used to train, validate and test the prediction model.

The training module in this phase is processed with the training subset and labels for those patterns. It introduces the MSVM model and stores it in the secondary storage device to load and perform test phases.

Testing the Churn analysis process predicts the possibilities on the given input dataset. The test set loads the test subset and the training module and then proceeds with the prediction method of MSVM. This method is a promising approach for predicting online datasets because MSVMs use a risk-minimization rule that contains the error.

After the prediction module, various performance parameters are used to calculate the performance of the proposed architecture. The computed performance metrics are used to build the comparison sets and validate the contribution of this enhancement.

V. SIMULATION RESULT ANALYSIS

This section describes the detailed dataset description, simulation tool, and performance metrics such as accuracy rate, AUC, etc.

A. Dataset Description

This proposed work has used "WA_Fn-UseC-Telco-Customer-Churn" dataset [14]. This proposed database defines

the raw information consisting 7043 rows ('users') and 21 columns (properties). Throughout the regression and categorization operations, characteristics 21 are employed as the target position. The churn columns are our target values.

B. Arithmetic Formula's

For the classification job with Optimized kernel MSVM, three parameters are utilized to measure the evaluation of the CCs (Churn Classification). They are AUC, Accuracy Rate (%), and RMSE rate. The area under the curve may be understood as an aggregate amount of classification evaluation to complete all expected categorization techniques. The accuracy rate may be expressed as;

$$Accuracy = 100 * \frac{\text{no.of correctly classified samples}}{\text{no.of all samples in the database}} \quad (1)$$

For the prediction and classification method with optimized kernel MSVM algorithm, RMSE (Root Mean Square Error Rate) is used to measure the calculation of the monthly charge. It may be expressed as;

$$RMSE = \sqrt{\frac{\sum_i^n (y1_i - y'1_i)^2}{n1}} \quad (2)$$

For the classification method with optimized kernel MSVM algorithm, the accuracy rate is the ratio of the correctly labeled churn values to the complete dataset. It may be defined as;

$$accuracy = \frac{tp+tn}{tp+fp+fn+tn} \quad (3)$$

Here, tn = True Negative, fp = False Positive, tp = True Positive, fn = False Negative, and n1 represents the number of all samples in the tele-comm dataset, $y1_i$ and $y'1_i$ are the intended and expected values, respectively. The database is divided into two modules to individually assess the approaches, such as testing and training modules. The training program is employed to the training purpose of the models, while the testing section is used to evaluate the models' efficiency. A 10-fold cross-validation methodology is acceptable including both CC (churn categorization) and forecasting activities to successfully perform the analysis. It indicates that the information is divided into subgroups with an equivalent amount of data in each. The residual set is also used to evaluate the strategy whereas alternative subgroups are used for training. This method is continued till all the subgroups have been analyzed separately and the training phase has been completed.

C. Implementation Results

This section describes the results of the churn analysis, and the research work has designed a script using the PYTHON language. This research work has worked on two different modules as the train and test module. The training module is created using the PYTHON language and is made. It will make the user interface for the man-to-machine interactions that the customer can easily click and see the output. The research will predict the test-set classes using an optimized kernel MSVM trained model and evaluate the accuracy rate, and AUC score shown in Fig. 2.

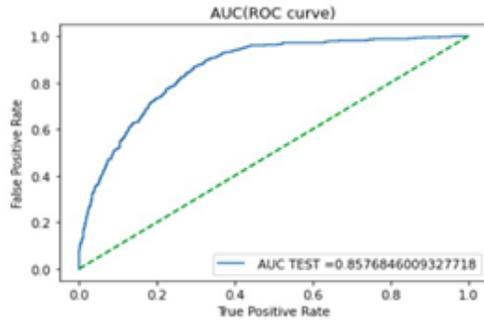


Fig. 2. AUC Score with the Proposed Model.

Generally, the classification and regression methods will perform better than the existing model. As the model utilized will be able to learn designs from the data, the proposed parameters are defined in Table III and shown in Fig. 2, 3 and 4 respectively. The table defines the performance metrics such as accuracy rate value of 91.05 per cent, rate value of 85.76 per cent, and RMSE Score value of 2.838. This proposed model will improve the performance metrics and reduce the error rates.

Initially, the proposed model will use train samples to construct a telecommunication sector customer churn analysis model with kernel PCA (principal component analysis) feature extraction. Table III defines the proposed parameters of various churn prediction methods. So, the proposed work will evaluate FE (feature extraction) using the KPCA model on all data to fetch the extract the feature values if train samples and test information, and now utilize the train samples which evaluate the optimized the dimensionality feature sets to reconstruct an OKMSVM (optimized kernel MSVM) model.

TABLE III. PROPOSED MODEL PERFORMANCE METRICS

Parameters	Optimized Kernel MSVM Classifier
Accuracy Rate (%)	91.05
AUC (%)	85.76
RSME score	2.838

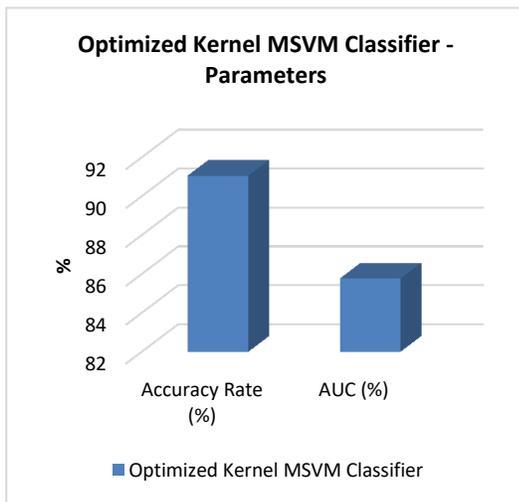


Fig. 3. Performance Metrics (Accuracy and AUC) with Optimized Kernel MSVM Classification Model.

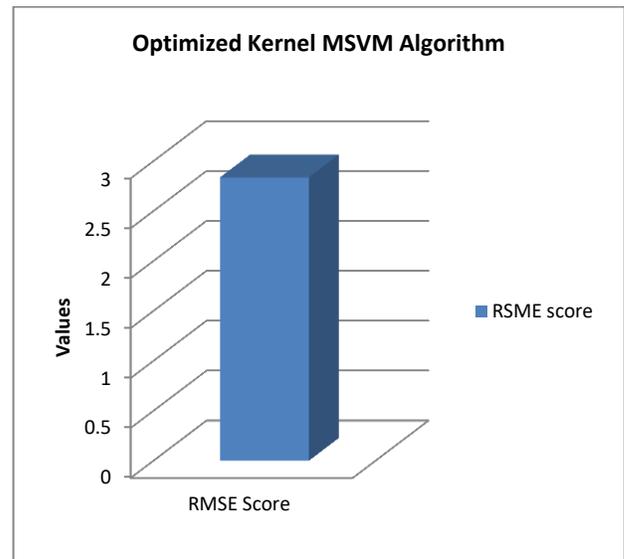


Fig. 4. Performance Metrics (RMSE) with Optimized Kernel MSVM Classification Model.

VI. CONCLUSION AND FUTURE WORK

This paper proposed a hybrid classification model named OKMVM model. The proposed model is based on a system for CCA (customer churn analysis) of the telecommunication sector. To find high-rate mathematical information and implement it, a feature extraction process is handled by the KPCA algorithm in the proposed architecture. Extracted parts are used to process with the Ant Lion optimizer. ALO is an optimization module of the proposed architecture that helps reduce the feature set's error probability. The more miniature error probability training set provides more accuracy in the trained model. It is an iterative process to find the best cost solutions for input feature patterns. Once all the cases are processed, the data element is processed with the initialization module of the prediction model. This module processes the optimized data elements and builds various subsets of the actual datasets. These subsets are used to train, validate and test the prediction model. The training module in this phase is processed with the training subset and labels for those patterns. It introduces the MSVM model and stores it in the secondary storage device to load and perform test phases. The test set loads the test subset and the training module and then proceeds with the prediction method of MSVM. The classification outcomes are maximum accuracy rate, AUC, and optimized root means square rate compared with existing models. In future, the researchers can design novel ML-based and SC (soft computing) techniques that are more reliable for enhancing performance metrics. Furthermore, a forecasting framework will be developed to determine individuals likely to churn. Approaches like regression analysis, decision trees, and neural networks will create multiple algorithms to create forecasting models. Also, generalization ability can be specified using various FS (feature selection) and classification techniques and their uses over other databases.

REFERENCES

[1] S, E., "A proposed churn prediction model." International Journal of Engineering Research and Applications 2, no. 4 (2012): 693-697.

- [2] V. Lazarov, and M Capota. "Churn prediction." *Bus. Anal. Course. TUM Comput. Sci* 33 (2007): 34.
- [3] B, Ionut, and G Todorean. "Churn prediction in the telecommunications sector using support vector machines." *Margin* 1 (2013): x1.
- [4] Zhao, M., Zeng, Q., Chang, M., Tong, Q, and Su, J. "A Prediction Model of Customer Churn considering Customer Value: An Empirical Research of Telecom Industry in China." *Discrete Dynamics in Nature and Society* 2021 (2021).
- [5] Bandara, W. M. C., Perera, A. S., and Alahakoon, D "Churn prediction methodologies in the telecommunications sector: A survey." In 2013 international conference on advances in I.C.T. for emerging regions (ICTer), pp. 172-176. IEEE, 2013.
- [6] Jain, N., Tomar, A., and Jana, P. K "A novel scheme for employee churn problem using multi-attribute decision making approach and machine learning." *Journal of Intelligent Information Systems* 56, no. 2 (2021): 279-302.
- [7] Sarac, F., Şeker, H., Lisowski, M., and Timothy, A "A Hybrid Two-Level Support Vector Machine-Based Method for Churn Analysis." In 2021 5th International Conference on Cloud and Big Data Computing (ICCBDC), pp. 77-81. 2021.
- [8] Lalwani, P., Mishra, M. K., Chadha, J. S., and Sethi, P. "Customer churn prediction system: a machine learning approach." *Computing* (2021): 1-24.
- [9] Bayrak, A. T., Aktaş, A. A., Susuz, O., & Tunali, O. "Churn prediction with sequential data using long short term memory." In 2020 4th International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT), pp. 1-4. IEEE, 2020.
- [10] Jain, H., Khunteta, A., and Srivastava, S. Churn prediction in telecommunication using logistic regression and logit boost. *Procedia Computer Science*, 167(2020), 101-112.
- [11] Ullah, I., Raza, B., Malik, A. K., Imran, M., Islam, S. U., & Kim, S. W. "A churn prediction model using random forest: analysis of machine learning techniques for churn prediction and factor identification in telecom sector." *IEEE Access* 7 (2019): 60134-60149.
- [12] Ahn, J., Hwang, J., Kim, D., Choi, H., & Kang, S "A Survey on Churn Analysis in Various Business Domains." *IEEE Access* 8 (2020): 220816-220839.
- [13] Townsend, A., and Nilakanta, S "Customer Churn: A Study of Factors Affecting Customer Churn using Machine Learning." (2019).
- [14] BlastChar, 2018. Telco customer churn. Kaggle. Available at: <https://www.kaggle.com/blastchar/telco-customer-churn> [Accessed January 31, 2022].
- [15] Yihui, Q., and Chiyu, Z. "Research of indicator system in customer churn prediction for telecom industry." In 2016 11th International Conference on Computer Science & Education (ICCSE), pp. 123-130. IEEE, 2016.
- [16] Acero-Charaña, Carlos, Erbert Osco-Mamani, and Tito Ale-Nieto. "Model for Predicting Customer Desertion of Telephony Service using Machine Learning", *International Journal of Advanced Computer Science and Applications* (IJACSA),12(3),2021.
- [17] Khedra, MM Abo, et al. "A Novel Framework for Mobile Telecom Network Analysis using Big Data Platform", *International Journal of Advanced Computer Science and Applications*(IJACSA),11(8),2020.
- [18] Jadhav, Rahul J., and Usharani T. Pawar. "Churn prediction in telecommunication using data mining technology." *International Journal of Advanced Computer Science and Applications*, (2011).
- [19] Alboukaey, Nadia, Ammar Joukhadar, and Nada Ghneim. "Dynamic behavior based churn prediction in mobile telecom." *Expert Systems with Applications* 162 (2020): 113779.
- [20] Seymen, Omer Faruk, Onur Dogan, and Abdulkadir Hiziroglu. "Customer Churn Prediction Using Deep Learning." *International Conference on Soft Computing and Pattern Recognition*. Springer, Cham, 2020.
- [21] Hu, X., Yang, Y., Chen, L., & Zhu, S. (2020, April). Research on a customer churn combination prediction model based on decision tree and neural network. In 2020 IEEE 5th International Conference on Cloud Computing and Big Data Analytics (ICCCBDA) (pp. 129-132). IEEE.
- [22] Ahmed, Ammar AQ, and D. Maheswari. "An enhanced ensemble classifier for telecom churn prediction using cost based uplift modelling." *International Journal of Information Technology* 11.2 (2019): 381-391.
- [23] Yu, Ruiyun, et al. "Particle classification optimization-based B.P. network for telecommunication customer churn prediction." *Neural Computing and Applications* 29.3 (2018): 707-720.
- [24] Abou el Kassem, Essam, et al. "Customer Churn Prediction Model and Identifying Features to Increase Customer Retention based on User Generated Content." *IJACSA International Journal of Advanced Computer Science and Applications* 11.5 (2020).
- [25] Khan, Yasser, et al. "Customers churn prediction using artificial neural networks (ANN) in telecom industry." *Editorial Preface From the Desk of Managing Editor* 10.9 (2019): 2019.
- [26] Butgereit, Laurie. "Big Data and Machine Learning for Forestalling Customer Churn Using Hybrid Software." 2020 Conference on Information Communications Technology and Society (ICTAS). IEEE, 2020.
- [27] Ewieda, Mahmoud et al., "Review of Data Mining Techniques for Detecting Churners in the Telecommunication Industry." *Future Computing and Informatics Journal* 6, no. 1 (2021): 1.

Study on Feature Engineering and Ensemble Learning for Student Academic Performance Prediction

Du Xiaoming¹

School of Economics and Management
Jiangsu University of Science and Technology
Zhenjiang, China

Chen Ying², Zhang Xiaofang³, Guo Yu⁴

Zhangjiagang Campus
Jiangsu University of Science and Technology
Suzhou, China

Abstract—Student academic performance prediction is one of the important works in the teaching management, which can realize accurate management, scientific teaching and personalized learning by mining important features affecting the academic performance and accurately predicting academic. Due to the subjectivity of feature extraction and the randomness of hyperparameters, the accuracy of academic performance prediction needs to be improved. Therefore, in order to improve the accuracy of prediction, an academic prediction method based on Feature Engineering and ensemble learning is proposed, which makes full use of the advantages of random forest in feature extraction and the ability of XGBoost in prediction. Firstly, the feature importance is calculated and ranked by using the random forest method, and the optimal feature subset combined with the forward search strategy. Secondly, the optimal feature subset is input into the XGBoost model for prediction. The sparrow search algorithm is used to optimize the XGBoost hyperparameters to further improve the accuracy of academic prediction. Finally, the performance of the proposed method is verified through the experiments of the public data set. The experimental results show that the academic prediction method designed is better than the single learner prediction method and other integrated learning prediction methods. The accuracy result jumps to 82.4%. It has good prediction performance and can provide support for teachers to teach according to students' aptitude.

Keywords—Academic performance prediction; feature engineering; ensemble learning; random forest; XGBoost

I. INTRODUCTION

With the application and popularization of emerging technologies such as internet of things, big data and artificial intelligence, the intellectualization of education has been developing rapidly, which promoting the reform of education and triggering the change of educational paradigm[1]. At present, many colleges have carried out intelligent campus construction and collected a large number of data generated by educational activities [2]. By machine learning technology, how to mine the mode and value contained in the data has become one of the urgent problems to be solved. Academic performance prediction refers to the use of relevant theories such as pedagogy, computer science and statistics, which analyzing the data generated in the process of students' learning and predicting their future academic performance [3]. Through academic prediction, managers can effectively manage the school, carry out academic early warning for students with academic risk in advance, establish a dynamic early warning

mechanism, and accurately guide students out of difficulties. Teachers can realize scientific teaching; predict students' academic status and teaching effect in advance, so as to purposefully optimize teaching activities, formulating differentiated teaching plans, meeting the needs of students at different levels, and truly teaching students according to their aptitude. Students can realize learning personalization; find out in advance which behaviors are beneficial to learning and which behaviors affect learning effect in advance. In this way, it could consolidate beneficial behaviors and break bad habits. Therefore, whether from the perspective of efficient school management, scientific teaching of teachers or personalized learning of students, to investigate the characteristics of affecting students' performance and create a high-accuracy prediction model has become an important research direction under the background of educational intelligence.

Currently, the research on the prediction of students' academic performance has some problems such as: subjectivity of feature extraction and poor prediction accuracy. This paper proposes an academic prediction model based on feature engineering and ensemble learning. Firstly, the out of bag estimation method of random forest is used to calculate and rank the feature importance, and the forward sequence method is used to search the optimal feature subset, which could deal with the randomness of the selection of random forest feature subset. Then the sparrow search optimization algorithm is used to adjust the hyperparameters of XGBoost to obtain the optimal combination of hyperparameters. Finally, the optimal feature subset is input into the optimized XGBoost model for academic prediction. The two kinds of ensemble learning methods are combined orderly, making full use of the advantages of random forest in feature extraction and the ability of XGBoost in prediction, which could enhance the generalization and effectiveness of academic prediction methods.

II. RELATED WORK

With the extensive use of data acquisition equipment, the collection of student data has expanded from single-mode learning data to multi-mode data, which including learning behavior, life behavior and psychological behavior [4]. The data show exponential development in volume and high dimensionality in characteristics. The essence of prediction is to find the mapping relationship between features and targets. Because the original feature set contains associated features and redundant features, it is often necessary to extract features in order to achieve better prediction effect. Too little feature

extraction will lead to "under fitting", which will affect the prediction accuracy. Too much feature extraction will lead to "over fitting", which will not only increase the calculation difficulty, but also reduce the prediction accuracy. In order to extract the optimal feature subset, some researchers manually select the feature subset through domain knowledge or expert experience. For example, Hu and others divide the features into static features and dynamic features, which takes students' basic information features as static features and students' behavior features (early rising behavior, borrowing behavior, etc.) as dynamic features, and predict learning performance according to the selected features [5]. Fan divides the characteristics into three types: tendency characteristics, human-computer interaction characteristics and interpersonal interaction characteristics. It is pointed out that the prediction ability of propensity characteristics is strong in the early stage of learning. With the progress of learning, the prediction ability of human-computer interaction characteristics and interpersonal interaction characteristics have gradually enhanced [6]. Li constructed a student behavior analysis model including five dimensions: Students' basic information, classroom learning, extracurricular learning, campus life and entertainment [7]. Some researchers use filtering or packaging based feature engineering methods to automatically select feature subsets. For example, based on filtered correlation analysis and information gain method, Chen and others calculate the Pearson correlation coefficient between features and scores, sorting them in descending order according to the results, and select the first 9 features from the 16 features as the main influence features [8]. However, it does not make the optimal selection of features. Through correlation analysis, information gain ratio and chi square analysis, Febro has selected 14 features from 29 original features to form the optimal subset, and verified that the prediction result of the feature subset is better than that of the whole feature [9]. Cao extracted the characteristics of students' regularity and preciseness from the campus life data. It is found that students' regularity is positively correlated with their grades, and preciseness is significantly correlated with their grades [10]. Wen proposed a hybrid feature selection method based on packaging. This method first generates candidate feature sets through scoring and sorting, and then uses heuristic methods to generate the final results [11]. However, the computational complexity of this method is exponential and needs longer running time. The above research screened the features, but did not consider the redundancy between features, and did not check the dimension of the optimal features.

In terms of prediction methods, machine learning methods have gradually replaced probability and statistics methods and gradually become the main research methods, including linear regression, logical regression, decision tree, support vector machine, neural network, integration method and deep learning [12-14]. Wu compared four different performance prediction methods: decision tree, Bayesian network, neural network and support vector machine, and found that the performance prediction model of Bayesian network has high accuracy and recall [15]. Liu used support vector machine to predict students' grades [16]. Wang used correlation analysis and regression analysis to study the predictive effect of big five personality traits and individual intelligence on academic

achievement [17]. Considering the influence of the spatial and temporal characteristics of students' behavior data, Du proposed a serial hybrid deep learning algorithm of CNN and LSTM to predict learners' performance [18]. Cao proposed LSTM depth neural network method to predict learning achievement [19]. Ding uses the methods of random forest, SVM, KNN, decision tree and naive Bayes to predict students' academic performance. The results show that the prediction performance of random forest algorithm is the best [20]. Yao proposed a multi task learning achievement prediction framework based on learning ranking algorithm. [21]. The above studies mostly use the single classifier method for prediction. It is found that the integrated learner has better performance and higher accuracy than the single learner. Ensemble learning methods include two types: boosting and bagging. Boosting methods include AdaBoost, GDBT and XGBoost etc. Some researchers have applied ensemble learning to many fields and achieved good results. Hao used XGBoost model to predict whether learners can complete the course and obtain certificates [22]. Xu used XGBoost model to automatically identify students' classroom behavior [23]. Cao uses XGBoost to predict the online short rent market price. The experimental results show that XGBoost is better than the integrated learning method of LightGBM and AdaBoost [24]. When using ensemble learning methods, the above researchers often use default hyper parameters or set hyper parameters based on experience. Because there are many kinds of hyperparameters, these methods often cannot obtain the optimal hyperparameters combination, which will affect the prediction accuracy.

III. METHODOLOGY

A. The Framework of Academic Prediction

This paper designs the academic prediction framework, as shown in Fig. 1. Academic prediction mainly includes four processes. First, academic data preprocessing. It consists of three parts: clean up, convert and normalize the data. Second, feature extraction. The random forest model is used to rank the importance of data features, and the optimal feature subset is extracted according to the forward search strategy. Third, model training. Train XGBoost model based on sparrow search optimization algorithm. Fourth, performance evaluation. The performance of the model is evaluated according to the evaluation metrics.

B. Feature Engineering

Feature engineering is an important link in the process of machine learning prediction. It can effectively remove the associated features and redundant features, and use the appropriate search strategy to extract the optimal feature combination, which is helpful to reduce the complexity and improve the accuracy of the prediction method. Random forest (RF) is an integrated learning method based on decision tree. Its embedded feature importance evaluation mechanism has the function of analyzing the correlation between features. RF has the advantages of simplicity and good robustness in feature extraction. RF belongs to bagging method. Samples are randomly selected from the original data for basic learner training. The unselected data is called out of bag (OOB), which can be used as a test set. The error predicted according to OOB

is called generalization error. For a feature, the generalization error is calculated after its eigenvalues are randomly disrupted. If the difference between the two generalization errors is small, it means that the feature is not important, otherwise it means that the feature is important. When calculating the feature importance of random forest, the out of bag data is used to calculate the generalization error before and after the disturbance of feature data, and the difference between the two generalization errors is calculated. The calculation steps of RF characteristic importance are as follows:

1) $D = \{(X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n)\}$ is the original data set $X_i \in R^T, Y_i \in R$, T represents the number of features and n represents the number of samples. m ($m < n$) samples are randomly selected from D in k times to generate k training sets, k OOB sets and k decision trees.

2) Calculate the generalization error e_t of OOB dataset corresponding to the t-th decision tree.

3) Keep other eigenvalues of OOB unchanged, randomly disrupt the order of eigenvalues of the i-th feature, and recalculate the generalization error e_t^i .

4) Repeat steps 2) - 3) to traverse the whole forest and calculate the importance of the i-th feature. As shown in formula (1):

$$\varepsilon^i = \frac{1}{k} \sum_{t=1}^k e_t - e_t^i \quad (1)$$

Through the above steps, the feature importance is calculated and sorted in descending order according to the importance. In order to extract the optimal feature subset, a forward search strategy is adopted. In the first round, the feature subset containing one feature is selected from the ordered feature set $F = \{f_1, f_2, \dots, f_t\}$. Obviously, the first feature f_1 is the selected feature subset of a single feature. In the i round, the first i features are formed into feature subsets, and their operation effects are compared with those of the first i-1 feature subsets. If the prediction accuracy is not as good as that of the first i-1 feature subset, the operation will be stopped, and the first i-1 feature subset is the best feature subset.

C. XGBoost Algorithm

XGBoost is an improved gradient boosting decision tree algorithm, which takes CART as the base learner and combines many base learners into high-performance integrated learners. For dataset $D = \{(\mathbf{x}_i, y_i)\} (i=1, 2, \dots, n)$, where \mathbf{x}_i represents the eigenvalue vector of sample i and y_i represents the label of sample i. Assuming that the XGBoost model contains T base learners, the predicted value of sample i is:

$$y_i = \sum_{t=1}^T f_t(\mathbf{x}_i), f_t \in F \quad (2)$$

F is all base learner spaces and $f(\mathbf{x})$ is the base learner function.

GBDT approximates the objective function by first-order Taylor expansion, while XGBoost approximates the objective function by second-order Taylor expansion to accelerate the convergence and improve the accuracy of the algorithm. In addition, in order to control the structural complexity of the model, XGBoost adds a regular term to the objective function to prevent the algorithm from over fitting and improve the generalization performance. The XGBoost objective function is:

$$obj(\theta) = \mathcal{L}(\theta) + \Omega(\theta) \quad (3)$$

$\mathcal{L}(\theta) = \sum_{i=1}^n (y_i - \hat{y}_i)$ is the error function.

$\Omega(\theta) = \sum_{t=1}^T \Omega(f_t)$ is a regular item.

In the process of XGBoost training, the next base learner is trained according to the residuals of the previous trained model to minimize the objective function. After much iteration, an integrated model with high accuracy is generated. The objective function at iteration t is:

$$obj^t = \sum_{i=1}^n ((y_i - y_i^{t-1}) - f_t(\mathbf{x}_i)) + \Omega(f_t) + C \quad (4)$$

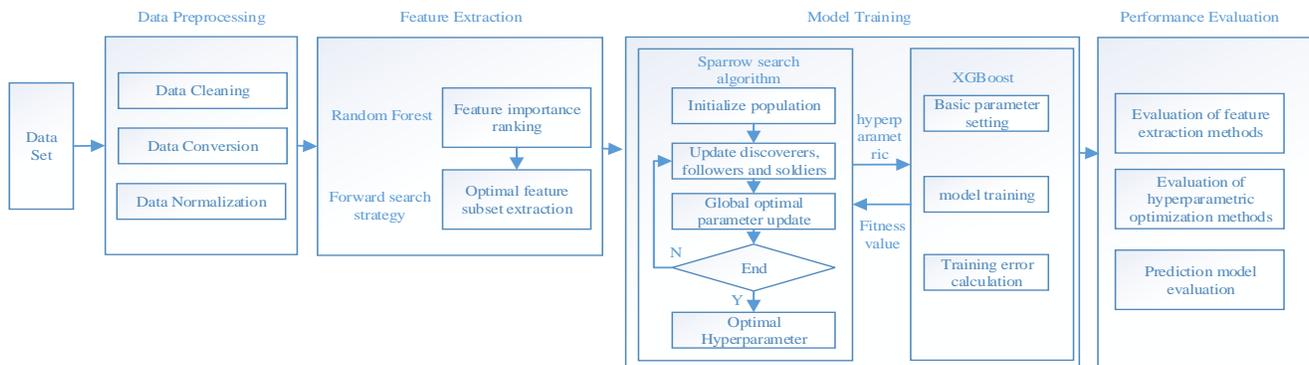


Fig. 1. Academic Prediction Framework.

f_t is a new basic learner function. C is constant. Carry out the second-order Taylor expansion of formula (4):

$$obj^t \approx \sum_{i=1}^n [L(y_i - y_i^{t-1}) + g_i f_t(\mathbf{x}_i) + \frac{1}{2} h_i f_t^2(\mathbf{x}_i)] + \Omega(f_t) + C \quad (5)$$

$g_i = \partial_{y_i} L(y_i, y^{t-1})$ is the first derivative.

$h_i = \partial_{y_i}^2 L(y_i, y^{t-1})$ is the second derivative. The complexity of XGBoost consists of the number of leaves and the structure of the tree, so the regularization term can be defined as:

$$\Omega(f_t) = \gamma T + \frac{1}{2} \lambda \sum_{j=1}^T w_{j(x)}^2 \quad (6)$$

γ and λ are constant. T is the number of leaf nodes. $w_{j(x)}$ is the real value of the leaf node to the sample. Therefore, formula (5) can be changed to:

$$obj^t \approx \sum_{i=1}^n [L(y_i - y_i^{t-1}) + g_i f_t(\mathbf{x}_i) + \frac{1}{2} h_i f_t^2(\mathbf{x}_i)] + \gamma T + \frac{1}{2} \lambda \sum_{j=1}^T w_{j(x)}^2 \quad (7)$$

The constant C does not affect the maximization of the objective function, so it can be omitted.

Define the sample set contained in the leaf node j :

$$I_j = \{i | \mathcal{Q}(\mathbf{x}_i) = j\} \quad (8)$$

According to formula (8), formula (7) is expressed as:

$$obj^t = \sum_{j=1}^T [(\sum_{i \in I_j} g_i) w_j + \frac{1}{2} (\sum_{i \in I_j} h_i + \lambda) w_j^2] + \gamma T \quad (9)$$

$$G_j = (\sum_{i \in I_j} g_i), H_j = (\sum_{i \in I_j} h_i)$$

$$obj^t = \sum_{j=1}^T [G_j w_j + \frac{1}{2} (H_j + \lambda) w_j^2] + \gamma T \quad (10)$$

The minimization of the objective function is transformed into the minimization of the quadratic function of w_j , and the optimal w_j^* is solved:

$$w_j^* = -\frac{G_j}{H_j + \lambda} \quad (11)$$

$$obj^t = \frac{1}{2} \sum_{j=1}^T \frac{G_j^2}{H_j + \lambda} + \gamma T \quad (12)$$

D. Sparrow Search Algorithm for Optimizing Hyperparameters

Although XGBoost improves GBDT, optimizes the convergence speed and improves the accuracy, the determination of XGBoost hyperparameters is still the key problem to improve its performance. The prediction effect is often not high when the hyperparameters are set according to experience. Therefore, it is necessary to set the hyperparameters through the optimization algorithm. Sparrow search algorithm (SSA) is a bionic intelligent optimization algorithm that simulates the foraging behavior and anti-predation behavior of sparrows [25]. Compared with other intelligent optimization algorithms, it has better global search ability, less iterations and high prediction accuracy. In the process of sparrow foraging, it is divided into discoverer, follower and police soldier. The discoverer updates his position according to the foraging rules and guides the population to forage. The follower obtains the food around the discoverer or competes for the food of other individuals and updates the position. When the sparrow group realizes the danger, it will carry out anti predation behavior and update the corresponding position. The discoverer update location rule is:

$$X_{i,j}^{t+1} = \begin{cases} X_{i,j}^t \exp(-\frac{i}{\alpha iter_{max}}), & R_2 < ST \\ X_{i,j}^t + Q, & R_2 \geq ST \end{cases} \quad (13)$$

$X_{i,j}^t$ represents the position of the i -th sparrow in the j -th dimension at the t -th iteration. $\alpha \in (0, 1]$ is a random number. $iter_{max}$ is the total number of iterations.

$R_2 \in [0, 1]$ is the warning value. $ST \in [0.5, 1]$ is the safety value. Q is a random number with normal distribution.

L is the matrix of $1 \times d$, whose values are all 1. $R_2 < ST$ indicates that the sparrow is in a safe state and can expand the foraging range. $R_2 \geq ST$ indicates that when a predator is found, all sparrows should fly away quickly.

The rule for followers to update the location is:

$$X_{i,j}^{t+1} = \begin{cases} Q \exp(-\frac{X_{worst}^t - X_{i,j}^t}{i^2}), & i > n/2 \\ X_p^{t+1} + |X_{i,j}^t - X_p^{t+1}| A^+ \cdot L, & i \leq n/2 \end{cases} \quad (14)$$

n represents the number of all sparrows. X_p and X_{worst} represent the position of the current discoverer and the worst position of all sparrows, respectively, L is the matrix of $1 \times d$, whose values are all 1. A^+ represents the pseudo inverse matrix. When $i > \frac{n}{2}$, it means that followers with low fitness value are difficult to capture food and need to fly to other places for feeding.

When aware of the danger, the sparrow updates the location rule as follows:

$$X_{i,j}^{t+1} = \begin{cases} X_{best}^t + \beta |X_{i,j}^t - X_{best}^t|, & f_i > f_g \\ X_{i,j}^t + k \left[\frac{X_{i,j}^t - X_{worst}^t}{(f_i - f_w) + \varepsilon} \right], & f_i = f_g \end{cases} \quad (15)$$

X_{best}^t is the best position among all sparrows at present. f_i , f_g and f_w represent the current sparrow fitness value, the worst fitness value among all sparrows and the best fitness value among all sparrows. β is the random value obeying (0,1) normal distribution. $k \in [-1, 1]$ is a random number. When $f_i = f_g$, it indicates that the sparrow is on the edge of danger and needs to be close to other sparrows.

IV. EXPERIMENTAL OF PREDICTION MODELS

The experimental environment of this paper is 64 bit Windows 7 operating system, the CPU is i5-3317u, the RAM is 4GB, the programming language is python, and the compilation environment is PyCharm.

A. Dominant Set and Data Preprocessing

This paper uses the score data set collected by the learning management system (LMS) of the University of Jordan. The data set contains 480 student records of 12 courses in 2 semesters, and each record includes a total of 16 features. These 16 characteristics are divided into four categories: demographic characteristics, knowledge background characteristics, parental behavior characteristics and learning behavior characteristics, as shown in Table I. It is divided according to 16 eigenvalue categories, including 5 binary features, 7 nominal features and 4 numerical features. The grades of each student's course are divided into three grades: L (0 to 69), M (70 to 89) and H (90 to 100). There are 127 students with L, 211 students with M and 142 students with H.

Data preprocessing is a very important step in machine learning. Standardized data processing can eliminate the impact of different data dimensions on prediction accuracy, and help to improve prediction performance while maintaining data distribution. It usually includes data cleaning (missing value processing), data conversion and data normalization. There is no missing value in the data set. Five binary feature data are transformed into {0,1}, and seven nominal features are mapped into quantitative feature values. Finally, all data are normalized by min-max method.

TABLE I. CLASSIFICATION OF DATA CHARACTERISTICS

Attribute type	Feature
Demographic Characteristics	Gender,Nationality, Place of birth,Parent responsible for student
Knowledge Background Characteristics	Educational Stages, Grade Levels, Section ID,Topic,Semester
Parental Behavior Characteristics	Parent Answering Survey,Parent School Satisfaction
Learning Behavior Characteristics	Raised hand, Visited resources, Viewing announcements,Discussion groups,Student Absence Days

B. Model Evaluation Metrics

The prediction results of this experiment are divided into three levels, which belong to multi classification problem. Accuracy and kappa coefficient are used as evaluation metrics. The value range of the two metrics is [0,1]. The larger the metric value, the better the prediction performance. The calculation formulas are:

$$Accuracy = \frac{d'}{Z} \quad (16)$$

$$Kappa = \frac{Accuracy - P_e}{1 - P_e} \quad (17)$$

$$P_e = \frac{d_1 \times d_1' + d_2 \times d_2' + d_3 \times d_3'}{Z \times Z} \quad (18)$$

Z is the total number of samples. d' is the number of samples with correct predictions. d_i' is the number of samples with correct prediction of type i . d_i is the total number of class i samples.

C. Feature Extraction

Based on the preprocessed data, the relative importance of features is calculated from equation (1) by using the random forest method. The results are shown in Fig. 2. Among the 16 features, the importance gap of each feature is obvious, and the five most important features belong to learning behavior features. It shows that students' learning behavior in class has a great impact on course performance. Too many or too few features will affect the prediction accuracy. In order to obtain the optimal feature subset, remove the unimportant features in turn according to the feature importance in Fig. 2, and calculate the Kappa index values of different feature subsets, as shown in Fig. 3. When the number of feature sets increases from 1 to 12, the Kappa index shows an increasing trend as a whole, reaches the maximum when the number of feature sets is 12, and shows a downward trend when the number of feature sets increases from 12 to 16. The main reason is that when the number of features is relatively small, the model training is insufficient, which affects the prediction accuracy. When the number of features is too large, the complexity of the model increases, resulting in over fitting training, which will also

reduce the prediction accuracy. Therefore, this paper selects the top 12 features of feature importance ranking for subsequent model prediction.

D. Comparison of Hyperparameters Optimization Methods

XGBoost parameters are divided into general parameters, lifter parameters and task parameters. General parameter setting is the overall function of the model, lifter parameter setting is the basic learner function, and task parameter setting is the optimization step. The parameters to be adjusted in this experiment and their adjustment range are shown in Table II.

In order to verify the efficiency of sparrow search algorithm in XGBoost hyperparametric optimization, comparative experiments are carried out by using manual experience method, grid search algorithm, random search algorithm and hyperparametric optimization method based on sparrow search algorithm. The experimental results are shown in Table III.

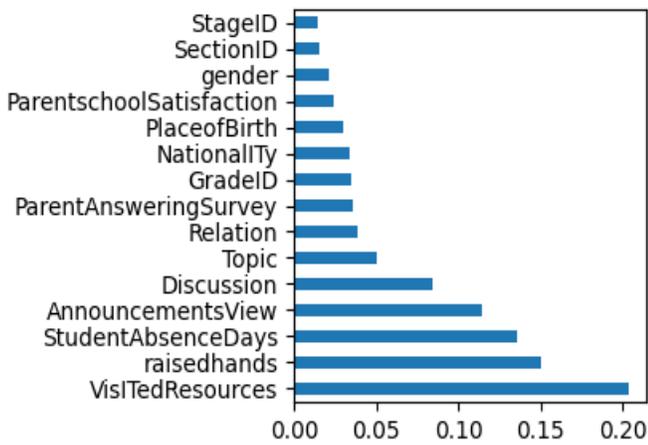


Fig. 2. Feature Importance.

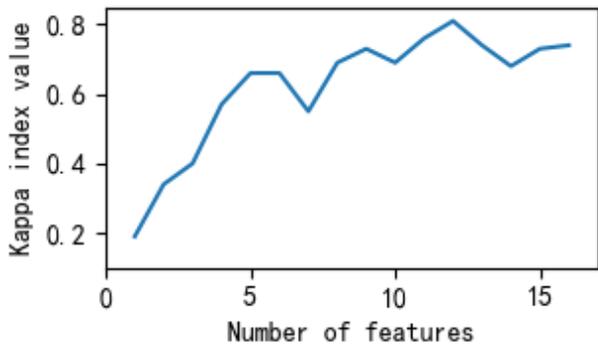


Fig. 3. Kappa Value.

TABLE II. HYPERPARAMETERS AND RANGES

hyperparameters	ranges	content
n_estimator	(100,1500)	Number of decision trees
learning_rate	(0.01,0.15)	Learning rate
max_depth	(1,10)	Maximum depth of decision tree
min_child_weight	(1,10)	Minimum weight of leaf node

TABLE III. KAPPA VALUES OF DIFFERENT METHODS

	manual experience method	grid search algorithm	random search algorithm	sparrow search algorithm
Kappa values	0.7755	0.8083	0.7924	0.8242

It can be found from Table III that the Kappa index value of the hyperparametric optimization method based on sparrow search algorithm is the highest, followed by the grid search algorithm, and the worst is the manual experience method. The results of sparrow search algorithm are 4.87%, 1.59% and 3.18% higher than manual empirical method, grid search algorithm and random search algorithm respectively in Kappa index. It is verified that different combinations of super parameters have a great impact on prediction performance. Theoretically, when the number and value range of super parameters to be optimized are large, there will be many combinations of super parameters. The grid search algorithm needs to exhaust the whole parameter combination space, and the time complexity is very high; Random search algorithm makes random sampling in a given space, which has fast search speed, but it is easy to miss some better parameter combinations; The sparrow search algorithm gradually obtains the optimal solution after iteration and updating the position, which has the characteristics of fast convergence and high accuracy.

E. Comparison of Different Machine Learning Methods

In order to verify the prediction performance of the method designed in this paper, it is compared with the mainstream single machine learning method and integrated learning method. Single machine learning methods include support vector machine (SVM), decision tree (DT) and logistic regression (LR). Integrated learning methods include gradient boosting decision tree (GBDT), XGBoost (XGB) with default value and the algorithm in this paper (SSA - XGB).

In order to avoid the contingency caused by random data division, five experiments are used to calculate the accuracy respectively, and the data set is divided into training set and test set according to 8:2. The experimental results are shown in Fig. 4, and the average accuracy is shown in Table IV.

TABLE IV. AVERAGE ACCURACY OF DIFFERENT ALGORITHMS

	SVM	DT	LR	GBDT	XGB	SSA- XGB
Accuracy	0.776	0.694	0.775	0.793	0.795	0.824

The experimental results show that the SSA- XGB achieves the best effect in performance prediction, followed by XGBoost with default value, and the worst is decision tree method.

Compared with XGB and GDBT, the SSA- XGB is 2.9% and 3.1% higher. The three integrated learning methods are better than the three single learner methods. Theoretically, the ensemble classifier is composed of multiple base learners. The prediction error of one base learner can be corrected by other base learners. The prediction error of a single classifier cannot be corrected. XGBoost algorithm improves the shortcomings of GDBT, such as the second-order Taylor expansion of the

objective function and the addition of regularization term. Compared with the default XGBoost algorithm, SSA-XGB shows that the optimization of hyperparameters is helpful to improve the prediction performance.

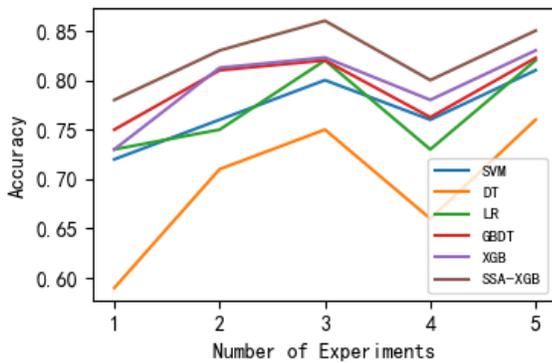


Fig. 4. Comparison of Accuracy of Different Algorithms.

V. CONCLUSION AND FUTURE WORK

The high dimension of academic data and the complex optimization of XGBoost hyperparameters are devoted to the paper research respectively. The important characteristics of academic data are extracted by random forest. The random forest algorithm using forward search strategy can effectively extract the feature subset and help to improve the prediction accuracy. The features which have a great impact on students' performance are mainly students' learning behavior characteristics, whose significance to 67 percent of all characteristics. Therefore, teachers should pay more attention to students' learning behavior status and help to improve students' academic performance. The adjustment of hyperparameters can improve the prediction performance of ensemble learning. Sparrow search optimization algorithm is more efficient than other methods of adjusting hyperparameters. Compared with other prediction methods, XGBoost prediction method has the perfect performance. The method designed in this paper enriches the methods of academic prediction in the field of educational data mining, and has a certain reference value for teachers to teach students according to their aptitude and students' personalized learning.

In future, the effects of different categories of guidance on students' academic performance can be studied respectively according to the characteristics of different categories. In addition, whether the combination of different features or the extraction of higher-order features is more conducive to academic prediction is also worth investing.

ACKNOWLEDGMENT

This research is supported by the National Educational Science Planning for the 13th Five Year Plan of China (Grant Number ECA180463).

REFERENCES

[1] Hang Hu, Yang Yang, "Pathways and strategies for deeper learning evaluation: a multi-modal data analysis," *Distance Education in China*, Vol. 9, PP.13-19, 2022.
[2] Song Dan, Liu DongBo, and Feng Xia, "Course Performance Prediction and Course Early Warning Research Based on Multi-source Data

Analysis," *Research in Higher Education of Engineering*, Vol. 1, PP. 189-194, 2020.
[3] Zhou Q, Mou C, and Yang D, "Research progress on educational data mining: A survey," *Journal of Software*, Vol. 26, No.11, PP. 3026-3042, 2015.
[4] Chen Kaiquan, Zhang Chunxue, et al., "Multi-modal Learning Analysis , Adaptive Feedback and Human-computer Coordination of Artificial Intelligence in Education," *Journal of distance education*, Vol. 37, No.5, PP. 24-34, 2019.
[5] Hang Hu, Shuang Du, and Jiarou Liang, "Towards a prediction model of learning performance: informed by learning behavior big data analytics," *Distance Education in China*, Vol.4, PP. 8-20, 2021.
[6] Yizhou Fan, and Qiong Wang, "Prediction of academic performance and risk: a review of literature on predicative indicators in learning analytics," *Distance Education in China*, Vol. 1, PP. 5-15, 2018.
[7] Li Y Z, and Hao Z, "Smart Campus Big Data Education Application Based on Students' Behavior Analysis Model," *China Educational Technology*, Vol. 7, PP. 33-38, 2018.
[8] Zijiang Chen, and Xiaoliang Zhu, "Research on Prediction Model of Online Learners' Academic Achievement Based on Educational Data Mining," *China Educational Technology*, Vol 12, PP. 75-81, 2017.
[9] Febro J D, "Utilizing Feature Selection in Identifying Predicting Factors of Student Retention," *International Journal of Advanced Computer Science and Applications*, Vol. 10, No. 9 , PP.269-274, 2019.
[10] Cao Y, Gao J, Lian D, et al., "Orderness Predicts Academic Performance: Behavioral Analysis on Campus Lifestyle," *Journal of The Royal Society Interface*, 2017.
[11] Wen Xiao, Ping Ji, and Juan Hu, "RnkHEU: A Hybrid Feature Selection Method for Predicting Students' Performance," *Scientific Programming*, PP.1-16, 2021.
[12] Rizvi S, Rienties B, and Khoja SA, "The role of demographics in online learning; a decision tree based approach," *Comput Educ*, Vol 137, PP. 32-47, 2019.
[13] Ashraf M, Zaman M, and Ahmed M, "An Intelligent prediction system for educational data mining based on ensemble and filtering approaches," *Procedia Comput Sci*, Vol 167, pp. 1471-1483, 2020.
[14] Huang AY, Lu OH, and Huang JC, "Predicting students' academic performance by using educational big data and learning analytics: evaluation of classification methods and learning logs," *Interactive Learn Environ*, Vol 28, No 2, PP.206-230, 2020.
[15] Qing Wu, and Ruguo Luo, "Predicting the Students' Performances and Reflecting the Teaching Strategies Based on the E-Learning Behaviors," *Modern Educational Technology*, Vol. 27, No. 6, PP 18-24, 2017.
[16] Liu B P, Fan T C, and Yang H, "Research on application of early warning of students' achievement based on data mining," *Journal of Sichuan University (Natural Science Edition)*, Vol. 56, No.2, PP. 267-272, 2019.
[17] Wang Jinyang, Wang Mengcheng, and Dai Xiaoyang, "Intelligence and personality predict high school students' academic performance: a big five Perspective," *Chinese Journal of Clinical Psychology*, Vol. 19, No. 6, PP. 824-826, 2011.
[18] Du Xiaoming, Ge Shilun, and Wang Nianxin, "Academic Prediction Based on CNN_LSTM Hybrid Neural Network," *Modern Educational Technology*, Vol.31, No.12, PP. 69-76, 2021.
[19] Cao Hongjiang, and XIE Jin, "LSTM-based Learning Achievement Prediction and Its Influencing Factors," *Journal of Beijing University of Posts and Telecommunications (Social Sciences Edition)*. Vol. 22, No. 6, PP. 90-100, 2020.
[20] Ding Xinfang, Nie Jing, and Zhang Bin, "Psychological factors and academic performance: a classified prediction model of machine learning," *Journal of Psychological Science*, Vol. 44, No. 2, PP. 330-339, 2021.
[21] Yao H , Lian D , Cao Y , et al. "Predicting Academic Performance for College Students: A Campus Behavior Perspective," *ACM transactions on intelligent systems*, Vol. 10, No.3, PP. 1-21, 2019.
[22] Hao Siyuan, and Xie Taifeng, "Study on the Application of Machine Learning in Predicting the Academic Completion Rate of MOOC

- Learners,” *Mathematics in Practice and Theory*, Vol.49, No.21, PP. 84-94, 2019.
- [23] Xu Jiazhen, Deng Wei, and Wei Yantao, “Automatic Recognition of Student’s Classroom Behaviors based on Human Skeleton Information Extraction,” *Modern Educational Technology*, Vol.30, No.5, PP. 108-113, 2020.
- [24] Cao Rui, Liao Bin, Li Min, “Predicting Prices and Analyzing Features of Online Short-Term Rentals Based on XGBoost,” *Data Analysis and Knowledge Discovery* Vol.5, No.6, PP. 51-65, 2021.
- [25] Xue J, and Shen B, “A novel swarm intelligence optimization approach : Sparrow search algorithm,” *Systems Science & Control Engineering*, Vol.8, No.1, PP .22-34, 2020.

Development of Hausa Acoustic Model for Speech Recognition

Umar Adam Ibrahim¹, Moussa Mahamat Boukar²
Computer Science, Nile University of Nigeria
Abuja, Nigeria

Muhammad Aliyu Suleiman³
Software Engineering, Nile University of Nigeria
Abuja, Nigeria

Abstract—Acoustic modeling is essential for enhancing the accuracy of voice recognition software. To build an automatic speech system and application for any language, building an acoustic model is essential. In this regard, this research is concerned with the development of the Hausa acoustic model for automatic speech recognition. The goal of this work is to design and develop an acoustic model for the Hausa language. This is done by creating a word-level phonemes dataset from the Hausa speech corpus database. Then implement a deep learning algorithm for acoustic modeling. The model was built using Convolutional Neural Network that achieved 83% accuracy. The developed model can be used as a foundation for the development and testing of the Hausa speech recognition system.

Keywords—Acoustic model; Hausa Phonemes; word level; CNN

I. INTRODUCTION

After several years of development and research, the accuracy of automatic speech recognition remains an important research issue. Many things influence the accuracy of a voice recognition application. Speaker and context variations are the most well-known. The use of acoustic modeling helps to improve accuracy. Any speech recognition system relies heavily on acoustic modeling.

The method of building statistic from voice waveform vectors is referred to as acoustic modeling for speech recognition. One of the most prevalent forms of acoustic model is the Hidden Markov Model [1, 2, 3]. Segmental models [4, 5, 6, 7, 8], super-segmental models such as hidden dynamic models [9], neural networks [10, 11], maximum entropy models [12], and hidden conditional random fields [13] are among the other auditory models.

Pronunciation modeling discusses fundamental units of speech sequences. These units include phonetics features. These phonetics features are utilized to present bigger speech units like phrases or words. To achieve noise robustness in speech recognition, acoustic modeling may also include the use of feedback [13].

The acoustic model describes the statistical features of sound occurrences in speech recognition. According to the acoustic model, the likelihood score $p(X|W)$ is calculated. Let's assume the acoustic model component representing an i th word W_i is Y_i , the $p(X|W) = p(X|Y_i)$ in an isolated-word speech recognition program with N -word vocabulary.

Hidden Markov Model (HMM), Support Vector Machine (SVM), Deep Neural Network (DNN), Artificial Neural Network (ANN), and Convolutional Neural Networks (CNNs) are some of the modeling techniques used [14].

The current work on acoustic model focused on international languages such English, French, Germany etc. Thus acoustic model developed for under-resource language are for Punjabi an India language, Amharic Ethiopian language and others [14].

The researcher implements a supervised Convolutional Neural Network. The Hausa acoustic model dataset was extracted from the Hausa Speech Corpus Database [15]. The researcher outlined the Hausa language alphabet and phonologies with examples.

The rest of the paper is organized as follows: Section II talked about Hausa phonology. Section III provides a review of acoustic modeling. Section IV presented the research methodology. Section VI described the implementation process. In Section V the result obtained was discussed and presented. Lastly, Section VII concludes this paper and mentions potential future work.

The major contribution of this paper is creating word-level phonemes, generating labels for the created phonemes, developing a Hausa acoustic model, testing and validating the developed model.

II. LITERATURE REVIEW

The Hausa language is a Chadic language spoken in West Africa. Hausa has two writing scripts. First, the Ajami writing script is written with the Arabic alphabet, and the Boko scriptwriting is written with the Latin alphabets.

This research is concerned with the Hausa Latin writing script. The Boko script has 36 Latin alphabet, which contain 31 consonants and 5 vowels. The language also has 23 phonetic sounds [16] as shown in Table I.

According to [17] Wurin/Gurbin furunci means a place of articulation. The consonant location in the vocal tract when a blockage occurs between active and passive articulators is known as the place of articulation. There are seven points of articulation for Hausa consonants which are:

Balebe meaning Bilabial: this is when the lower lip brushes up against or touches the upper lip.

Bahanke meaning Alveolar: when the tip of the tongue meets or touches the alveolar ridge.

Nade-harshe means Retroflex: when the tip of the tongue makes contact with the back of the alveolar ridge.

Dan Bayan Hanka meaning Post-alveolar: when the blade of the tongue comes close to or contacts the back of the alveolar ridge.

Bagande meaning Palatal

Bahande meaning Velar: is when the back of the tongue rubs on the velum or soft palate.

TABLE I. HAUSA ALPHABETS

A	B	B	C	D	D
E	F	Fy	G	Gw	Gy
H	I	J	K	K	Kw
Ky	Kw	Ky	L	M	N
O	R	S	Sh	T	Ts
U	W	Y	Y	Z	'
Consonants					
B	B	C	D	D	F
Fy	G	Gw	Gy	H	J
K	K	Kw	Ky	Kw	Ky
L	M	N	R	S	Sh
T	Ts				
Vowels					
A	E	I	O	U	

TABLE II. HAUSA PHONOLOGY

s/n	Phonology	Phonetic	Sample Word
1	Balebe=> Bilabial	/b/	“baka”
		/b̄/	“barawo”
		/m/	“malam”
		/f/	“fata”
2	Bahanke => Alveolar	/d/	“dankali”
		/n/	“talata”
		/r/	“nama”
		/z/	“bara”
		/s/	“zakka”
		/l/	“lada”
		/ts/	“tsawa”
3	Nade-harshe => Retroflex	/t̄/	“ruwa”
		/d̄/	“daki”
4	Dan Bayan Hanka =>Post-Alveolar	/ʃ/	“shara”
		/dʒ/	“jarida”
		/ɟ/	“canji”
5	Bagande =>Palatal	/j/	“yara”
6	Bahande =>Velar	/g/	“gado”
		/k/	“kaza”
		/k̄/	“kasa”
		/w/	“wasa”
7	Hamza => Glottal	/h/	“hannu”
		/ʔ/	“sa’a”

Hamza means Glottal: this refers to the closure or constriction of the glottis. Table II shows Hausa phonology, phonetic and sample word according to [17].

Acoustic Modeling (AM) is the first and crucial step in developing a speech recognition platform. The acoustic model generates a link between linguistic and acoustic units. Most of the computations performed in acoustic modeling are because statistical representation and feature extraction affect speech recognition development.

The extracted features are distributed based on a particular sound. This acoustic modeling is done to build a link between the structure of the linguistic model unit and the extracted features.

Different feature extraction methods, like voice production mechanism, and human perception were reported in [18, 19].

The selection of classification algorithms is also a crucial phase in the development of an acoustic model. Many studies on acoustic modeling using various classification algorithms have been published [20]. HMM, ANNs, DNNs, and Sequence to sequence acoustic modeling are some of the classification approaches used by researchers.

AM is associated with a variety of concepts. It necessitates knowledge of acoustic phonetics, microphone, and environment variability issues, gender variations, and dialectal variances. Furthermore, thorough training is required to determine the link between language units and auditory observation [21]. AM is also linked to pronunciation modeling, as well as speaker, environment, and context variability and modeling [22].

III. METHODOLOGY

A generalized acoustic modeling system includes raw data collection, preprocessing, feature extraction, and acoustic modeling. Fig. 1 illustrated the researchers’ acoustic modeling. The pipeline is divided into two components word-level phonology segmentation and word-level phonology labeling. The preprocessing block consists of word-level segmentation and word-level phonology labeling. Furthermore, feature extraction was implemented as the second block in the pipeline. The second block comprises phonemes and labels matched together for modeling purpose. The output of the feature extraction block was fed into the training block for acoustic modeling of the phonology utterances. Finally, a word-level Phonemes based acoustic model was implemented for this research. The word-level acoustic dataset was extracted from the Hausa Speech Corpus database [15].

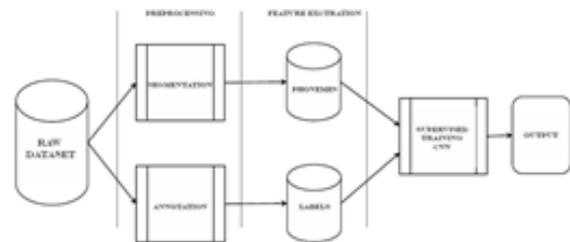


Fig. 1. Hausa Acoustic Model Pipeline.

A. Data Distribution

Five thousand word-level phonemes were extracted from the Hausa Speech Corpus database [15]. Male and female speakers read these words. Table III shows the number of words per literature and the gender of the reader.

TABLE III. DISTRIBUTION OF WORD-LEVEL PHONE AND READER GENDER

S/N	Literature	Words	Gender
1	Gani Gareka	500	Female
2	Iliya dan Maikarfi	700	Female
3	Iki Magayi	700	Male
4	Wani Gari	500	Male
5	Komai Nisa Dare	500	Male
6	Koya Da kanka	700	Female
7	Magana Jarice	700	Male
8	Shehu Umar	700	Male
Total		5000	

The final audio dataset consists of Five thousand phonetically compact word-level phonemes. The phonetically compact words were divided into training, testing, and validation set. The dataset distribution is shown in Table IV.

TABLE IV. SPEECH DATASET

S/N	Division	Percentage	words
1	Training	70%	3500
2	Testing	20%	1000
3	Validation	10%	500

B. Model

Various categorization techniques have been created for audio modeling. HMM, and ANNs are, however, the most extensively utilized algorithms [16]. This work was implemented using deep learning acoustic modeling algorithm. The researchers developed the Hausa acoustic word-level models by implementing a Convolutional neural network (CNN).

IV. IMPLEMENTATION

A. Hardware and Software

The system was built on top of Google Collaborator (Colab) with the following hardware spec:

- CPU: Intel(R) Xeon(R) @ 2.30GHz.
- Disk: Size 79GB, Used-40GB, Available-39GB.
- Memory: Total:~13GB, Free:~10GB, Available:~12GB.

For the software requirement, all necessary modules and dependencies were installed and imported into Colab. Such as OS, Pathlib, MAplotlib, Numpy, Seaborn, Tensorflow, Keras, and Ipython display. The audio dataset is stored in eight different folders corresponding to each literature named: gani, iliya, jiki, wani, koya, and shehu. The audio clips were extracted and shuffle into a list. The dataset were divided into training, test, and validation set as 70:20:10 ratios, respectively.

B. Preprocessing

The dataset was processed at this phase by creating decoded tensors for the waveforms and labels. Each wav file contains time-series data that is sampled at a specific rate. The amplitude of the audio signal at any given time is represented by each sample. The researchers' WAV acoustic dataset files had amplitude values ranging from -32,768 to 32,767. A 16-bit system was employed. This dataset has a sample rate of 16 kHz.

The shape of the tensor returned by "ft.audio.decode_wav" is the [sample, channels]. The Hausa acoustic dataset contains mono recordings.

Three functions are defined: the first transform the dataset's WAV audio files into tensors. The second is the method that generates labels for each file based on its parent directories. Finally, there is the "get waveform and label" auxiliary function, which ties everything together.

The audio filename is the input and the output is a tuple with audio and label tensors, ready for supervised learning. The audio waveforms were plotted as shown in Fig. 2.

C. Waveforms to Spectrogram

The time domain is used to depict the waveforms in the collection. The waveforms are then converted to spectrograms. The converted spectrogram reveals the frequency variations over time which can be depicted as 2D images. Further, the Short-Time Fourier Transform (STFT) is used to convert from time-domain signals to time-frequency-domain signals. The spectrogram images are then fed into the neural network, which is used to train the model.

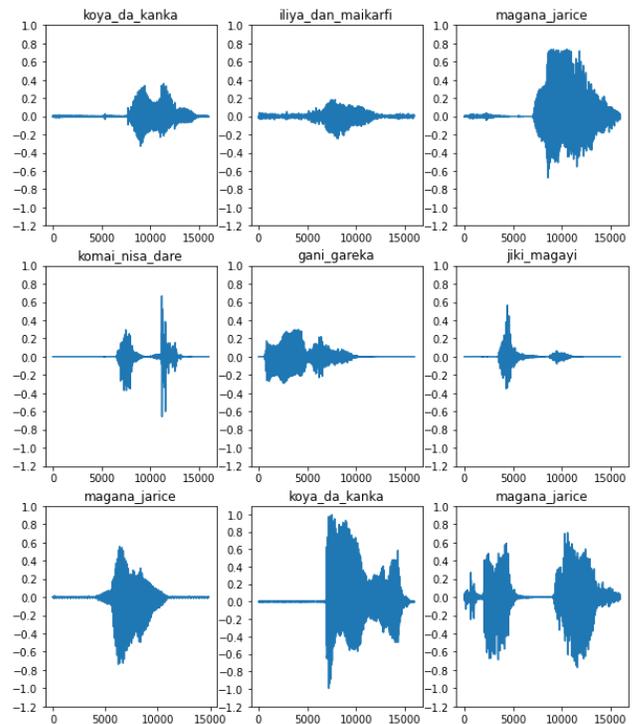


Fig. 2. Audio Waveforms.

Furthermore, a function for converting waveforms to spectrograms was created. The waveforms must be the same length for the spectrograms to have equal dimensions when converted. This can be accomplished by simple zero-padding audio snippets that are less than one second in length using (ft.zeros). In addition, a frame length and frame step options for tf.signal.stft are selected so that the resulting spectrogram "image" is almost square. The STFT generates a complex number of the array that represents magnitude and phase. In this situation, the researcher utilized the magnitude, which can be achieved by running ft.abs on the tf.signal.stft result.

Next, the data was explored by printing the shape of one sample tensorized waveform and the corresponding spectrogram. A function was defined to display a spectrogram. Lastly, sample audio was plotted displaying waveform over time and the corresponding spectrogram (frequencies over time) as shown in Fig. 3.

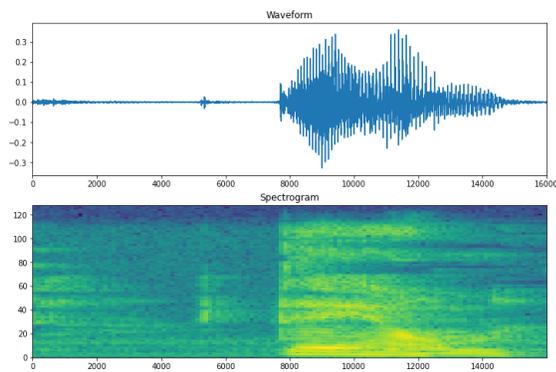


Fig. 3. Waveform and Spectrogram.

A function was then defined to transform the waveform dataset into spectrograms and their corresponding labels as integers. Map function was implemented across the dataset elements. Finally, the spectrograms for different samples of the dataset were examined as shown in Fig. 4.

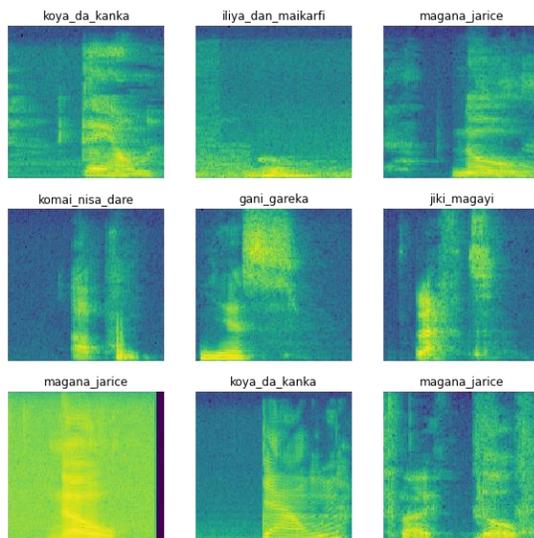


Fig. 4. Samples of Spectrogram.

D. Training the Model

Preprocessing was done on the dataset. Batch training and validation sets were implemented for the model training. Dataset.cache and Dataset.prefetch operations were performed to reduce read latency while training the model.

Since the audio files have transformed into spectrogram images. For modeling a simple Convolution Neural Network was implemented. Keras preprocessing layers were also implemented such as resizing layers to down sample the input to enable the model to train faster. A normalization layer was implemented to normalize each pixel in the image based on its means and standard deviation. The Keras model was configured with Adam optimizer and the cross-entropy loss. The model was trained on 10 epochs for demonstration purposes. The model summary is shown in Fig. 5.

```

Input shape: (124, 129, 1)
Model: "sequential"

```

Layer (type)	Output Shape	Param #
resizing (Resizing)	(None, 32, 32, 1)	0
normalization (Normalization)	(None, 32, 32, 1)	3
conv2d (Conv2D)	(None, 30, 30, 32)	320
conv2d_1 (Conv2D)	(None, 28, 28, 64)	18496
max_pooling2d (MaxPooling2D)	(None, 14, 14, 64)	0
dropout (Dropout)	(None, 14, 14, 64)	0
flatten (Flatten)	(None, 12544)	0
dense (Dense)	(None, 128)	1605760
dropout_1 (Dropout)	(None, 128)	0
dense_1 (Dense)	(None, 8)	1032

```

Total params: 1,625,611
Trainable params: 1,625,608
Non-trainable params: 3

```

Fig. 5. Model Summary.

For better analysis, the data was run on several epochs. However, 50 epochs gave better accuracy. The output of the model is shown in Fig. 7.

```

Epoch 1/50
55/55 [=====] - 11s 195ms/step - loss: 0.2411 - accuracy: 0.8200 - val_loss: 0.5089 - val_accuracy: 0.8200
Epoch 2/50
55/55 [=====] - 11s 194ms/step - loss: 0.2169 - accuracy: 0.8231 - val_loss: 0.5097 - val_accuracy: 0.8200
Epoch 3/50
55/55 [=====] - 11s 195ms/step - loss: 0.2206 - accuracy: 0.8100 - val_loss: 0.5663 - val_accuracy: 0.8220
Epoch 4/50
55/55 [=====] - 11s 195ms/step - loss: 0.1959 - accuracy: 0.8337 - val_loss: 0.6023 - val_accuracy: 0.8300
Epoch 5/50
55/55 [=====] - 11s 190ms/step - loss: 0.1933 - accuracy: 0.8331 - val_loss: 0.6209 - val_accuracy: 0.8200
Epoch 5: early stopping

```

Fig. 6. Model Output.

V. RESULT

After running the model with 50 epochs, the summary and output of the model as shown in Fig. 5 and 6. Finally, the training and validation loss curves were plotted to see how the model progressed over time as shown in Fig. 7.

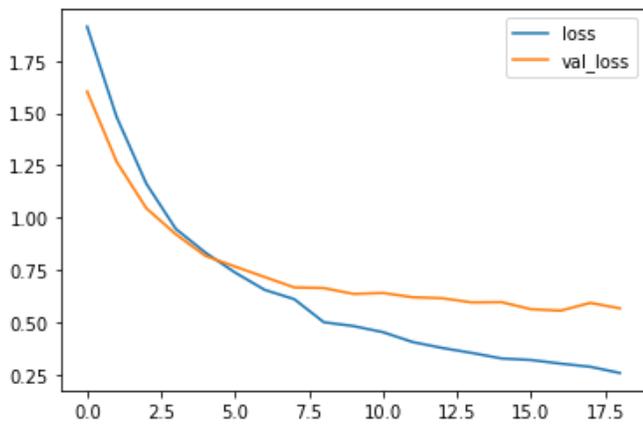


Fig. 7. Train-validation Loss Curves.

To evaluate the model’s performance, the model was run on the test set. The accuracy of the test was 83%.

To see how successfully the model classified each auditory word in the test set, a confusion matrix was plotted as displayed in Fig. 8.

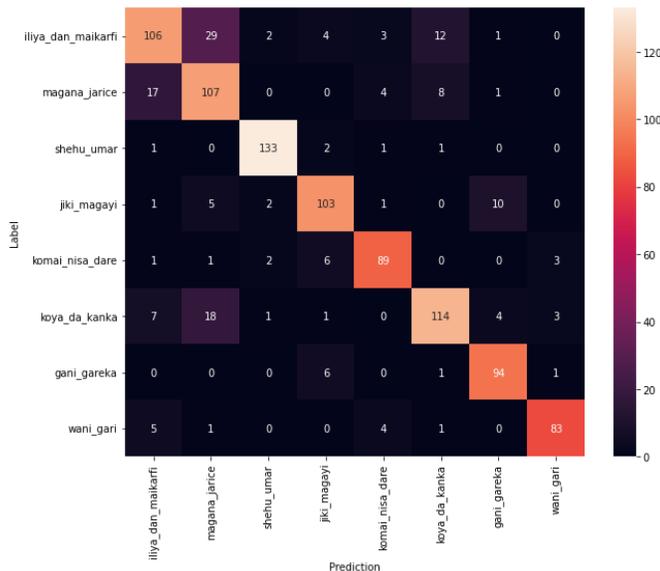


Fig. 8. Confusion Matrix.

VI. CONCLUSION

This article started with an introduction to the acoustic model. Stated the previous and current techniques used for the development of an acoustic model for speech recognition and then introduces Hausa phonology, its examples, and alphabets. The researchers created word-level phonemes from the Hausa Speech Corpus database.

With the rise of deep learning algorithms for acoustic model development, the researchers implemented CNN base acoustic model development for the Hausa language. Hausa is a low-resourced and under-resource language. The goal is to create an acoustic model for the Hausa language.

The created model can be used for the development of Hausa speech recognition system. The outputs suggest that the

model recognizes Hausa phonetic with 83% accuracy. The researchers’ future work is to develop a language model for the Hausa language. The acoustic model and language model would be linked to developing an automatic speech recognition system for the Hausa language.

REFERENCES

- [1] L. Baum, “An inequality and associated maximization technique occurring in statistical estimation for probabilistic functions of Markov process,” *Inequality III*, 1972, pp. 1–8.
- [2] J. Baker, “Stochastic modelling for automatic speech recognition in D.R Reddy,” In D.R Reddy, (ed), *Speech Recognition*, Academic Press, New York, 1975.
- [3] F. Jelinek, “A fast sequential decoding algorithm using a stack,” in *IBM journal of Research and Development*, vol. 13, pp. 675–685, 1969.
- [4] A. Poritz, “Hidden Markov Models: A guided tour,” in *Proceedings of the International conference on acoustic, Speech and Signal processing*, Vol. 1, pp. 1-4 1998.
- [5] L. Deng, “A stochastic model of speech incorporating hierarchical nonstationary,” *IEEE Transaction on speech and audio processing*, Vol. 1(4), pp.417-475, 1993.
- [6] L. Deng, M. Aksamanovic, X. Sun, and C. Wu, “Speech recognition using hidden markov models with polynomial regression function as nonstationary states,” *IEEE Transaction on speech and audio processing*, vol. 2, pp. 507–520, 2004.
- [7] M. Ostendorf, V. Digalaskis, and J. Rohlicek, “From HMMs to segment models: A unified view of stochastic modelling for speech recognition,” *IEEE Transaction on speech and audio processing*, vol. 4, pp. 360–378, 1996.
- [8] J. Glass, “A probabilistic framework for segment-based speech recognition,” in M. Russell and J. Bilmes(eds), *New computational paradigms for acoustic modelling in speech recognition computer, speech and language (special issue)*, 17(2-3), pp. 137–155, 2003.
- [9] L. Deng, D. Yu, and A. Acero, “Structured speech modelling,” *IEEE Transaction on speech and audio processing*, (special issue on rich transcriptin), vol. 14(5), pp. 1492–1504, 2006.
- [10] R. Lippma, “An introduction to computing with neural nets,” *IEEE ASSP Magazine*, 4(2), pp. 4–22, 1987.
- [11] N. Morgan, et. al, “Pushing the envelop-aside,” *IEEE signal processing magazine*, pp. 81–88, 2005.
- [12] Y. Goa, and J. Kuo, “Maximum entropy direct models for speech recognition,” *IEEE Transaction on speech and audio processing*, vol. 14(3), pp. 873–881, 2006.
- [13] A. Gunawardana, and W. Bryne, “Discriminative speaker adaptation with conditional maximum likelihood linear regression,” in *proceedings of the EUROSPEECH*, aalborg, Denmark, 2001.
- [14] S. Bhatt, A. Jain, and A. Dev, “Acoustic modelling in speech recognition: A system review,” *Internation Journal of Advance computer science and applications*, vol. 11, No. 4, 2020.
- [15] U.A. Ibrahim, M.M. Boukar, and M.A Suleiman, “Development of Hausa dataset a baseline for speech recognition,” *Dat in brief*, 40, 2022.
- [16] <https://wisc.pb.unizin.org>, “Hausa Alphabet” [online]. Available:<https://wisc.pb.unizin.org/lctresources/chapter/hausa-alphabet> [Accessed: 20-Nov-2021].
- [17] www.amsoshi.com, “Phonology 1(Wurin/Gurbin Furuci(Place of Articulation))” [online]. Available: <https://www.amsoshi.com/2020/02/alh-203-hausa-phonology-1-wuringurbin.html>. [Accessed: 20-Nov-2021].
- [18] A. Dempster, N. Laird, and D. Rubin, “Maximum likelihood from incomplete data via the EM algorithm,” *Journal of the Royal Statistical Society*, vol 39(1), pp. 1-21, 1977.
- [19] H. Hermansky, “Perceptual linear predictive (PLP) Analysis of speech,” *J. Acoust. Soc. Am.*, vol. 87, no. 4, pp. 1738-1752, 1990, doi:10.1121/1.39923.
- [20] S. Bhatt, A. Dev, and A. Jain, “Confusion analysis in phoneme based speech recognition in Hindi,” *J. Ambient Intell. Humaniz. Comput.*, no. 0123456789, 202, doi: 10.1007/s12652-020-01703-x.

- [21] M. J. F. Gales, S. Watanabe, and E. Fosler-Lussier, "Structured Discriminative models for speech recognition: An overview," *IEEE Signal Process. Mag.*, vol. 29, no. 6, pp. 70-81, Nov 2012. Doi: 10.1109/MSP.2012.2207140.
- [22] R. K. Aggarwal and M. Dave, "Acoustic modeling problem for automatic speech recognition system: Advances and refinements(part II)," *Int. J. Speech Technol.*, vol. 14, no. 4, pp. 309-320, 2011, doi:10.1007/s10772-011-9106-4

New Method for 4D Reconstruction of Medical Images

Lamyae MIARA, Said BENOMAR ELMDEGHRI, Mohammed Ouçamah CHERKAOUI MALKI

Department of Computer Science, University Sidi Mohammed Ben Abdellah
Faculty of Sciences Dhar El Mahraz, Fez, Morocco

Abstract—This paper proposes a new optimized method that is fast in rendering for 4D reconstruction from 2D medical images of human anatomy permitting their real-time refined visualization. This method uses the 3D reconstruction algorithm based on contour matching of medical image sequences and on the tessellation of recent GPU. In our framework, the construction of the low-resolution mesh that is based on contour extraction allows to create a 3D mesh without any ambiguity and exactly matches the real shape of the human anatomy. Such preliminary result is of great interest, since it permits to lead to other valuable realizations such as reducing the computation burden of basic meshes and displacement vectors. Moreover, one can achieve a very low storage memory, as well as one can ease the fast real-time 4D visualization with a high desired resolution. Hence, it is then straight forward that this study can contribute to easing the diagnosis and detection in real-time of human organs in motion damage and deterioration. Especially, 4D visualization technology that is still under development is highly important and needed for assessing some dangerously evaluative diseases, as in the case of lung diseases.

Keywords—2D medical image; 4D reconstruction; contour matching; recent GPU; 3D mesh

I. INTRODUCTION

From health and care perspectives, 4D visualization is the most effective method for assessing and preventing deadly diseases of human organs in motion. For instance, 4D visualization technology of lungs damaged can shed new light on the real possible dangers and eventual damage and complications on the patient health. As a real practical consequence, the use of computational methods in lung analysis may allow us to better understand and visualize what we cannot obtain from static 2D images, such as the respiratory behavior of muscles. To achieve this, the 4D reconstruction is an important and critical requirement.

More importantly, not to say that 4D reconstruction of the human organs in motion would be useful for visualization the correspondent mechanism, especially before surgery. For example, in patients suspected of having pulmonary diseases, the ability to visualize in 4D and observe the oxygen uptake capacity of the lungs would be useful for clinical diagnosis. So, a complete 4D reconstruction of the lungs may allow physicians to better understand the respiratory process (inspiration and expiration). Other disease states in which 4D modeling can potentially be useful for clinical decision making include pulmonary nodules, pneumothorax, and chronic diseases [1] characterized by irreversible decrease in bronchial caliber.

Unfortunately, the 4D reconstruction is a computationally intensive job, and therefore doctors use specific workstations [2].

In this paper, we propose a powerful and optimized algorithm to reconstruct a 4D image from segmented contours of 2D MRI images, to be able to do the 4D reconstruction in home computers.

Our proposed framework has the following advantages:

- Topological changes in 2D contours are automatically processed in both spatial and temporal dimensions.
- The result is a distortion-free mesh that corresponds to a good approximation that matches exactly the real human organs anatomy.
- The use of the tessellation of recent graphics cards makes the real-time display very fast.
- Memory usage is highly optimized by considering the large number of details in the 2D images.

The remainder of this paper is organized as follows. In section 2, an overview on the state of the art is presented and the limitations of some reported works are discussed. In section 3, our methodology is clearly exposed regarding to the previously reported studies. Section 4 presents some technical explanations and discussions on the proposed method and its obtained results. Finally, some necessary concluding remarks are provided in section 5.

II. RELATED WORK

Research studies on 4D reconstruction algorithms often focuses on both, improving the quality of 4D visualization from standard acquisition data and its fast execution. Among these algorithms is the Feldkamp-Davis-Kress (FDK) algorithm [3], which is a filtered back-projection algorithm widely used for 3D and 4D image reconstructions from cone-beam projections measured with a circular orbit of the x-ray source. There is also the Mc Kinnon-Bates (MKB) algorithm [4] in which a time-averaged 3D prior image is first reconstructed. It is then projected at the same angles as the original projection data, creating time-averaged re-projection that are next subtracted from the original (non-blurred) projections to create the well-known motion-coded differential projections. Such differential projections are reconstructed into PC differential images that are added to the well-sampled 3D image before creating a much higher quality 4D image.

Other distinguished algorithm is the ROOSTER algorithm [5] that iteratively computes volumes that minimize the sum of the least squares difference between simulated volume projections, real measured projections, and a spatiotemporal "total variation" term that favors distinct homogeneous regions with sharp edges. In addition, we also mention the Motion Compensated [6] FDK (MCFDK) algorithm [7] in which motion was estimated from 4DCT planning volumes by Deformable Image Registration (DIR) [8] between respiratory phase volumes using a B-spline method from the Elastix toolkit [9]. In addition, the 4DCT DIR produces Deformation Vector Fields (DVF) that are used as inputs to the MCFDK reconstruction. Note that the MCFDK algorithm can be interpreted as a variant of the FDK method where back-projection is performed along curved trajectories to account for motion. And lastly the Motion Compensated (MCMKB) algorithm which is an extension of [10], where 4DFDK reconstructions are used for DVF estimation. This study considers a much more under sampled acquisition than [10], which may explain why we were not able to produce convergent DVF estimates from the 4DFDK reconstructions.

Note that for the algorithms that use surface rendering, there is the classical method that uses interpolation between 3D images of to give a 4D result. These algorithms are effective alternatives for 4D reconstruction of medical images without the need for powerful machines, but they produce ghost artifacts that do not represent the exact anatomy of human organs and not achieve the determination of accurate motion trajectories. Another negative fact of using mentioned algorithms, is that it cannot lead to an instantaneous precise information on a particular organ and even to precise measurements on the organ to be reconstructed. In contrast, our proposed approach is based on the 4D surface rendering that allows a good visualization of the geometry and a precise topological shape associated to the three-dimensional structures of human organs and their spatial-temporal relationships. Moreover, such adapted procedure makes it possible for comparing the reconstructed data with the original data. For these reasons, it is stressed out that the surface rendering method eases the achievement of a fast real-time 4D visualization, thanks to the use of certain units of the graphic card originally conceived for 3D video games.

To recap, the main goal of this paper is mainly to propose a new algorithmic solution that is faster to compute based on accelerating rendering method and other efficient computational techniques. Moreover, the proposed methodology can achieve a high visual resolution by avoiding heavy drawbacks on memory storage and acquisition. Indeed, such improvement handles efficiently the limitations of the previously discussed algorithms that have been reported and recognized in the literature.

III. METHODOLOGY

This section describes the new method developed in this paper for an optimized and fast 4D visualization from standard 2D medical images. This method consists first of the 3D

reconstruction of a basic mesh from the 2D images of the C_m slices at time T_0 and the extraction of the lost data, in the form of displacement vectors, which will be stored in a displacement map. Thus, we perform the extraction of the displacement vectors of the 2D slices of each sequence from time T_1 to time T_{N-1} based on the mesh at time T_0 , i.e., without constructing the base mesh of times T_1 to T_{N-1} . In special cases, it is found that the number of contours at time T_i changes in time T_{i+1} , which makes it necessary to add additional parts in the base mesh. As a result, a high-resolution visualization can be achieved at the GPU map level by combining the different parts of the base mesh and the set of displacement maps.

The reminder of our contribution is as follows. In the first part given by section 3.1, we determine the input of our algorithm and explain the basic idea behind its conception. Next, the second and third and fourth parts considered in section 3.2 and 3.3 and 3.4 describe in more details the first steps of the proposed algorithm consisting of contours extraction and displacement maps. Some technical details are provided in section 3.5 to show how one can map the computed mesh at time T_i to the other at time T_{i+1} . Finally, it is shown how a high resolution 4D visualization can be handled in the last Section 3.6.

A. Overview

Our algorithm is based on a sequence of 2D medical images identified at successive time periods from time T_0 to time T_{N-1} . These images are subjected to a segmentation process to extract the target organ to be reconstructed and visualized in real-time. For this purpose, we apply the contour extraction algorithm on each slice C_k of each sequence S at each time T_i . In this way, we build the base mesh of the sequence at time T_0 and perform the extraction design of the corresponding displacement map, as well as for the other sequences (see Fig. 1). It is stressed out that, in the general, one may have a base mesh composed of other several bases mesh. The 4D reconstruction is then done with a single base mesh and $N+1$ displacement maps, which speeds up the rendering and optimizes memory usage.

In what follows, the illustration in Fig. 2 depicts the necessary steps to be executed to get the desired results for a fast and clear 4D visualization.

B. Extraction Des Contours

Based on the segmented slices, we need to extract the contours of the object to be reconstructed. For the detection of its contours, several methods can be possibly used. Such methods are grouped into several distinct classes. For instance, one can use those based on nonlinear filtering such as the median filter, or more recently the one in [11]. Other methods of interest are high-pass filtering, such as Prewitt, Sobel, and Canny detectors [12], or such that the multi-scale analysis developed with the wavelet theory [13] [14]; or the one based on the rare redundant dictionary approximation [15].

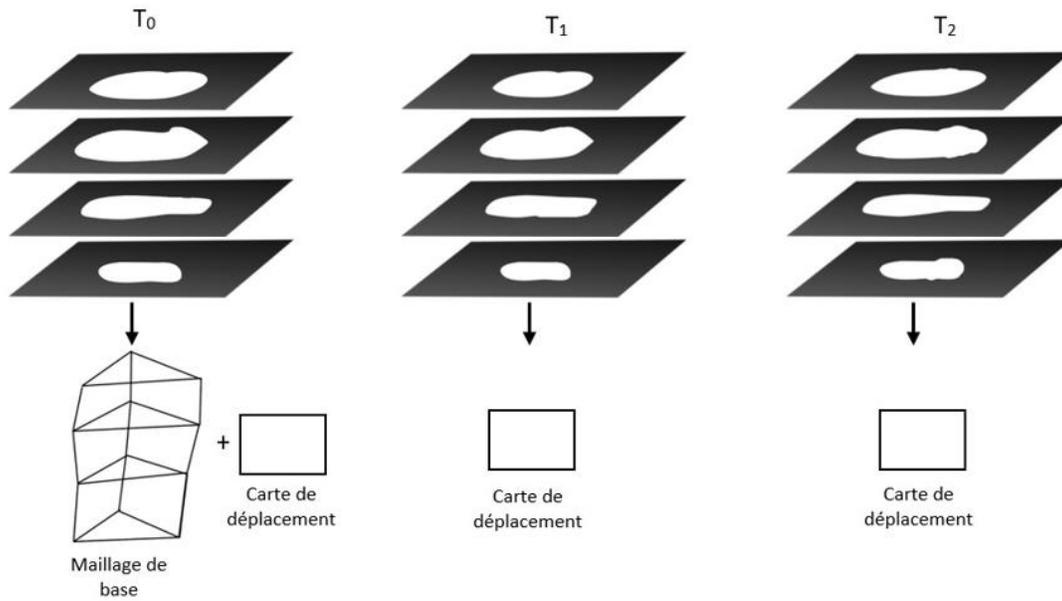


Fig. 1. Construction of Base Mesh and Displacement Map from a 2D Segmented Image Sequence.

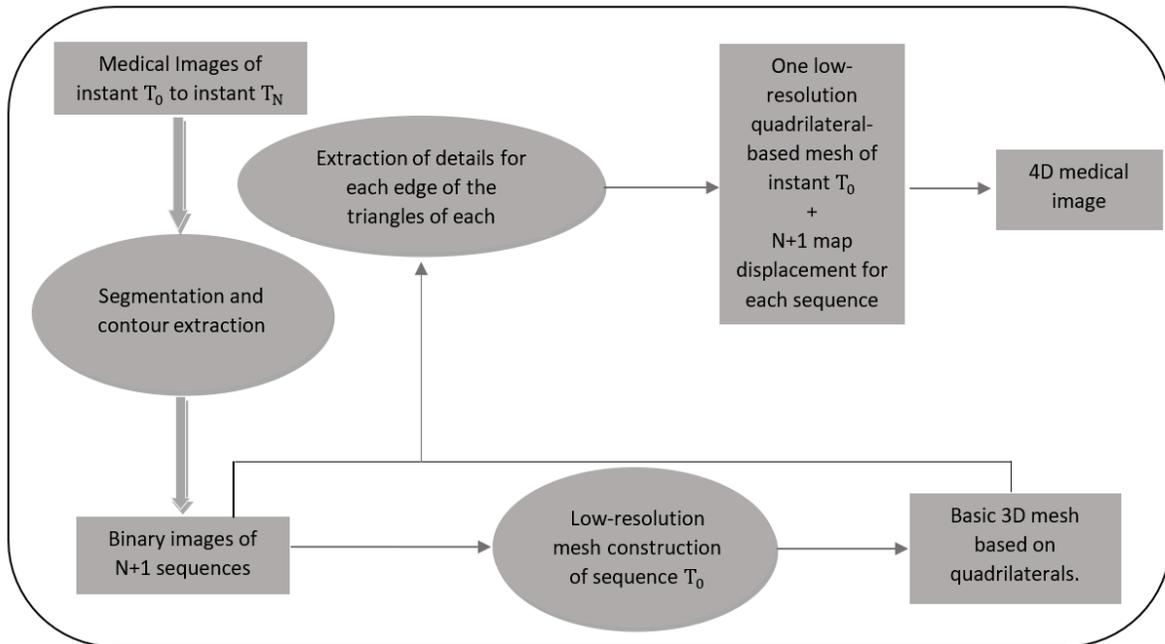


Fig. 2. Diagram Representing the Different Steps of our Algorithm

Note that the key role of the previously mentioned methods of contour traversal consists in defining the abrupt changes of pixel intensity. In fact, the real rising issue to overcome is that we should obtain a chained list of contour points, respecting a fixed order. So, the strategy we used is to extract a point from this contour, and then we must extract the other points in an ordered way following the path of these contours as it has been performed in [16].

To construct the displacement map, it is emphasized that the extraction of the contours shall be carried out for all the slices of all the sequences at the time periods T_i (see Fig. 3).

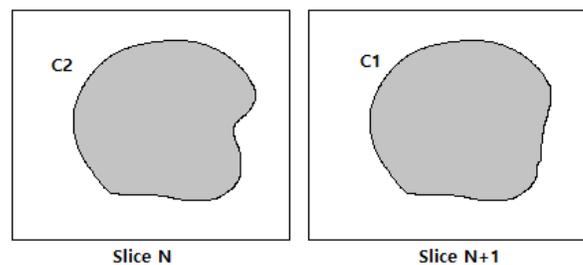


Fig. 3. Two Contours C_1 and C_2 of Two Successive Slices: the N Slice and the N+1 Slice at Time t .

C. Construction of the basic Mesh

The basic mesh is the first structure to be reconstructed without refined details. Note that in our design approach, the mesh construction consists in mapping the contours represented by triangles of each contour of each slice n with those of the slice $n+1$. As a matter of fact, drawing faces from the contours is not at all an easy task because this depends on the resolution of the following three hard problems [17]:

- Matching problem: How to connect the contours in slice n with a contour in slice $n+1$?
- Tiling problem: How to connect the points of contour C_n in slice n with the points of contour C_{n+1} in slice $n+1$?
- Connection problem: How to divide the contour C_i in slice n that corresponds to the contours C'_{n+1} and C''_{n+1} in slice $n+1$?

The illustration shown in Fig. 4 highlights the possibilities for solving the matching problem associated to the case when the number of contours in adjacent slices is not the same (the contours can be split or merged).

It is worth mentioning that for the determination of the relationship between the contours of two consecutive slices (see Fig. 5), we have used the correspondence factor method, see for instance [16].

D. Displacement Vectors

As it has been revealed in previous illustrations, the first constructed mesh is a basic mesh which does not reflect the true shape of the anatomy in question. It is then necessary to add to it the lost details in the construction of the base mesh during the real-time visualization. For this task, we shall then extract the displacement vectors using the available contour data and the edges of the quadrilaterals obtained after joining the triangles that represent the contours [18].

The strategy used to extract the displacement vectors is based on the discretization of the concerned contours (see Fig. 6). This approach has a great interest for two essential issues. On one hand, the generation of such discretized displacement vectors allows more precision for the construction of the real shape of the organ anatomy (see Fig. 7). On the second hand, through the discretization of the anatomy contour with a fixed point, this makes it easy to incorporate the suitable level of details in all polygons of the base mesh.

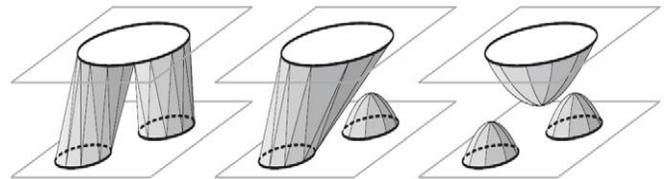


Fig. 4. Three Possible Solutions to the Topology Change Matching Problem.

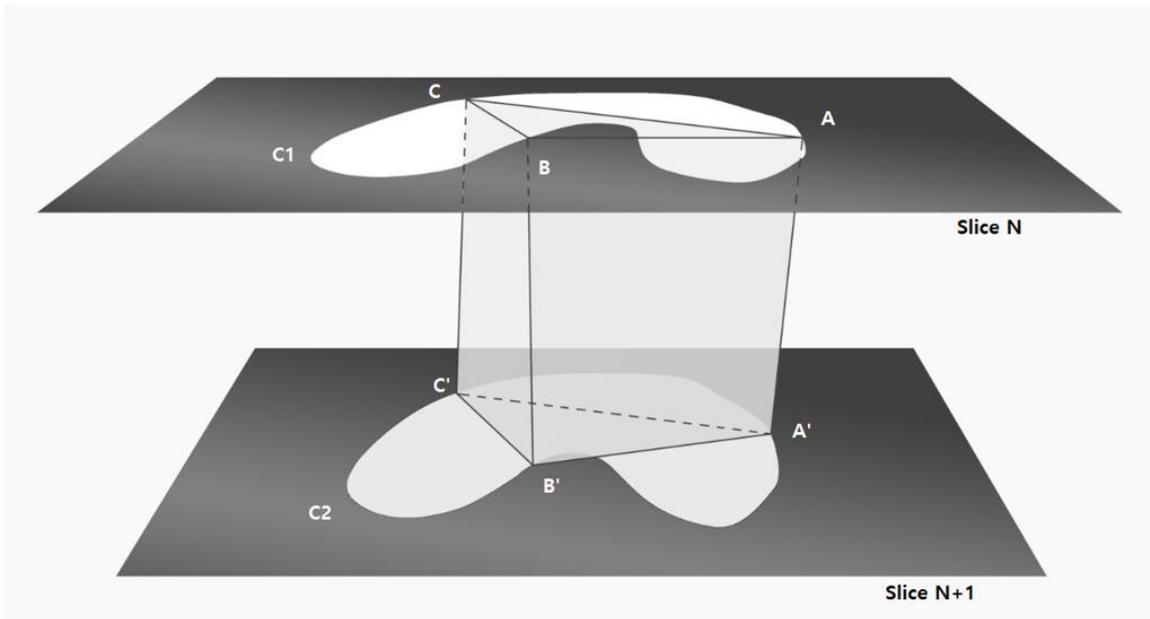


Fig. 5. Base Mesh Construction for Two Contours of Two Consecutive Slices.

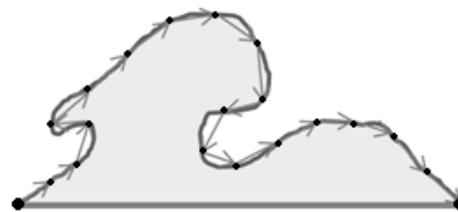


Fig. 6. Extraction of Contours with Displacement Vectors.

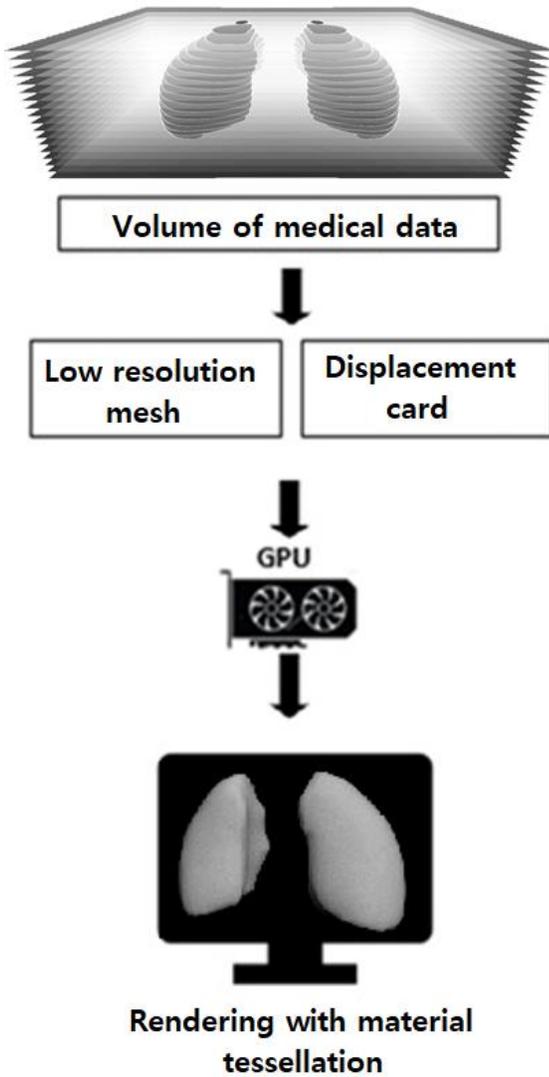


Fig. 7. Rendering with Hardware Tessellation to Generate the High-resolution Mesh.

Indeed, the key role of these constructed displacement vectors is that they can automatically generate the displacement map representing an image determined by its moving distances and directions from the points on the contour surface during the real-time display.

E. Reconstruction of the 4D Mesh

- Creation of sub-meshes and displacement vectors

Here, we are now able to characterize the 4D construction of the mesh. At the beginning of such construction, we start then from an already defined set of n sequences of slices from an initial time T_0 to a final time T_{n-1} . For each sequence we have m slices from C_0 to C_{m-1} . These slices are already segmented to extract only the object to be reconstructed (see Fig. 8).

This set of slices will be divided into $m-2$ sub-assemblies, each sub-assembly is represented by 2 consecutive slices of the n sequences. In particular, the first sub-assembly 0 will be represented by slice 0 and slice 1 of all the sequences.

Next, each subset must undergo some additional treatments as described in what follows:

For each sequence i we need to determine the different correspondences existing between the contours of 2 consecutive slices [16]. Contours that do not have a correspondent in the other slice will be neglected.

In Fig. 9, we have a correspondence between contour 1 and 2 of slice 1 and contour 1 and 2 of slice 2 respectively; contour 3 of slice 1 does not correspond to any contour in slice 2, so it will be deleted (see Fig. 10). Each correspondence gives us a base sub-mesh and displacement vectors that will be added to this mesh at time T_i .

The next step is to find for each sub-mesh its correspondent in the following defined sequences, for which we say that the correspondence of a sub-mesh of the sequence i to the sub-mesh of the sequence $i+1$ is satisfied if the contours of these two sub-meshes correspond to each other at each slice and with the same number (see Fig. 11).

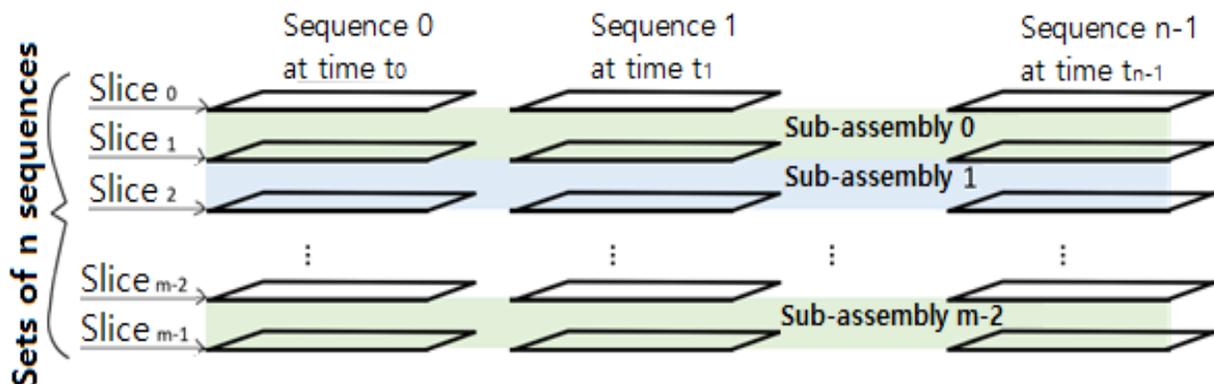


Fig. 8. The Input to this Algorithm is in the Form of a Sequence of 2D Slices for each Instant.

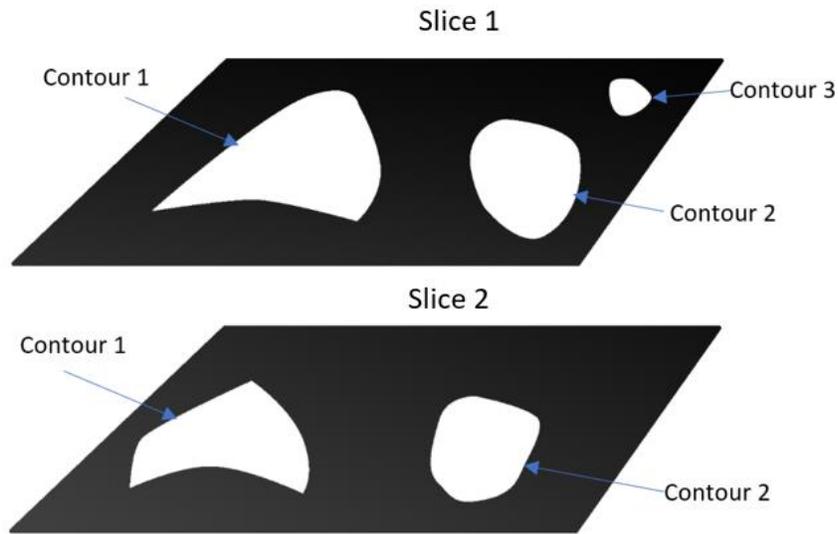


Fig. 9. Correspondence between the Contours of a Sequence.

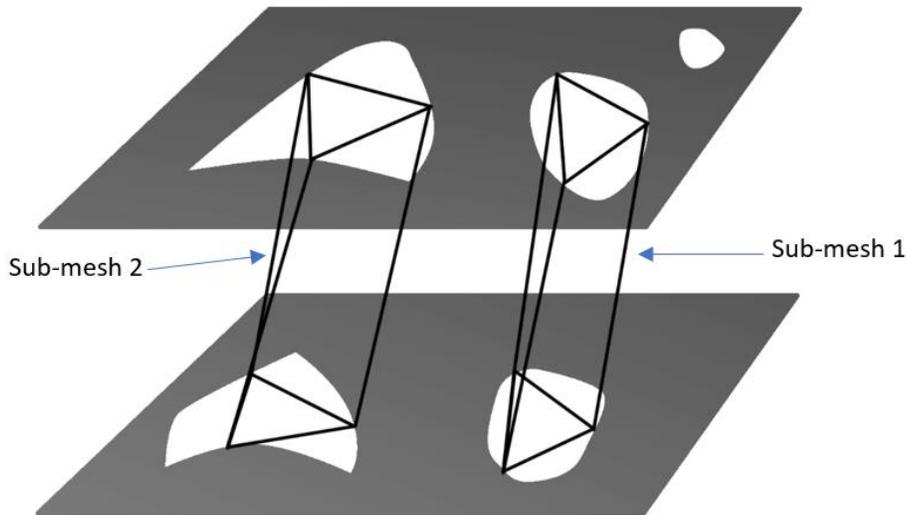


Fig. 10. Representation of the Sub-meshes which Represent the Correspondences between the Contours of the same Sequence.

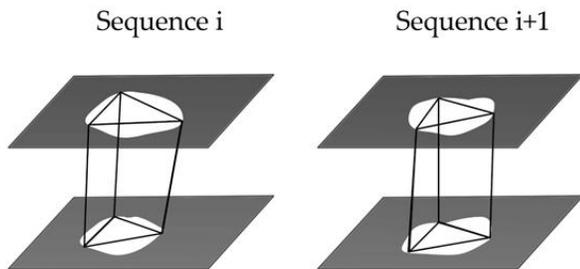


Fig. 11. The Sub-mesh of Sequence i Corresponds to the Sub-mesh of Sequence $i+1$.

If we fail to find the correspondence of a sub-mesh of a sequence i in the following sequence, the appearance of a new sub-mesh is obligatory, and it will be from the instant T_{i+1} (see Fig. 12).

Note that we can also have a new sub-mesh in a sequence i if we fail to find a correspondence between its contours and the contours of the sequence of the instant T_{i-1} (see Fig. 13).

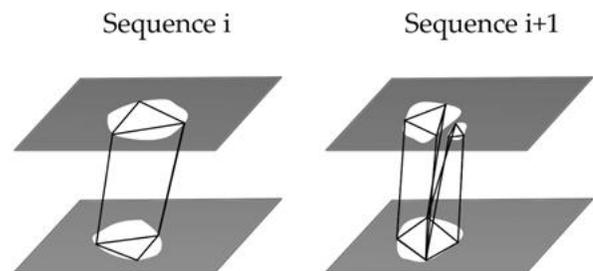


Fig. 12. No Correspondence between These Two Sub-meshes (the Number of Contours in Sequence i is different from the Number of Contours in Sequence $i+1$).

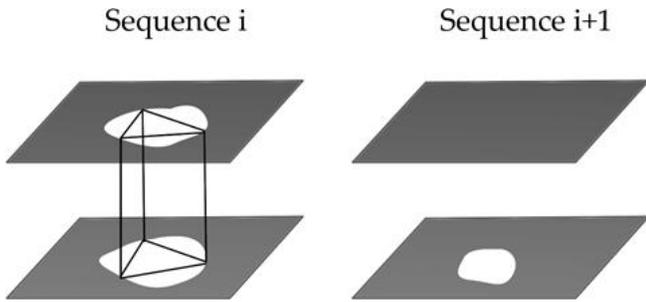


Fig. 13. This Sub-mesh does not have a Correspondent in the following Sequence.

Finally, after applying this additional treatment to all the subassemblies, we obtained series of sub-meshes, each sub-mesh has the following information:

- Start sequence number (S_d).
- End sequence number (S_f).
- Displacement vectors for each edge in the different sequences from the first sequence to the final sequence.

Note that each sequence has its own displacement vectors that refine the corresponding base sub-mesh. Another fact shown the above table in Fig. 14, that sequences from 0 to i, sequences from i+1 to j, sequences from j+1 to k, as well as sequences from k+1 to n-1 do not share the same basic sub-mesh in a sub-assembly.

- Construction of the 4D mesh and interpolation

Following the previous development for the obtained results from N given sequences, we have by now all the necessary ingredients to form 3D meshes for all sequences in the form of a single compound base mesh and N displacement maps. Thus, the next treatment is the visualization of the dynamic 3D mesh using all the available data and the gathered

information. However, the success of this crucial procedure results from the number of given sequences and the required quality of the visualization in terms of display smoothness. In fact, if the number of sequences is not sufficient or the required quality is very high; an extra interpolation is then obligated to perform for additional renderings to maintain the desired display quality between each consecutive sequences. Especially, the interpolation will be applied to the displacement vectors. In the normal case, where there is a correspondence between the meshes of sequence i and those of sequence i+1, the position of these interpolated meshes between the two sequences is deduced by creating intermediate displacement maps based on the already computed displacement maps in sequence i and sequence i+1. In the opposite case of no existing correspondence between the meshes associated to consecutive sequences (see Fig. 15); then, the calculation of the intermediate mesh can be carried out after finding the suitable correspondence between the contours.

In this case, for example, where there is no correspondence between the mesh of sequence i and the mesh of sequence i+1, we must create a new sub-mesh at time i.

This sub-mesh will have the sequence i as the starting sequence. Thus, its elements must correspond to the elements of the sub-mesh of sequence i+1. To do so, we will use the correspondence factor between the contours of slice j of time i and the contours of slice j of time i+1 [16], to create a correspondence between the contours of the same slice, the correspondence direction, and the correspondence percentage.

According to the calculated correspondence factor of the contour C_1 with the contours C_2 and C_3 of the sequence i+1, we find that the green part of the contour C_1 corresponds to the contour C_3 (see Fig. 16), and the orange part corresponds to the contour C_2 . Thus, we can build the sub-mesh below which will be the sub-mesh used from the instant i+1:

	Seq. 0	Seq. 1	...	Seq. i	Seq. i+1	...	Seq. j	Seq. j+1	...	Seq. k	Seq. k+1	...	Seq. n-1
S.-M. ₁	Start			end									
S.-M. ₂					start		end						
S.-M. ₃								start		end			
S.-M. ₄											Start		End

These sequences use the same basic sub-mesh for a sub-assembly i

Fig. 14. This Table shows that each Sub-mesh has a Start and an end Number.

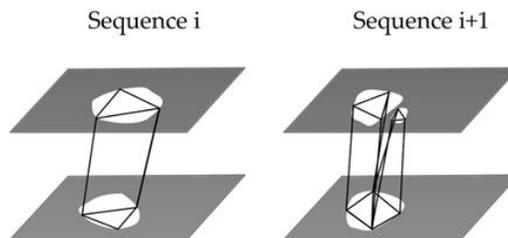


Fig. 15. No Correspondence between the Meshes of Two Consecutive Sequences.

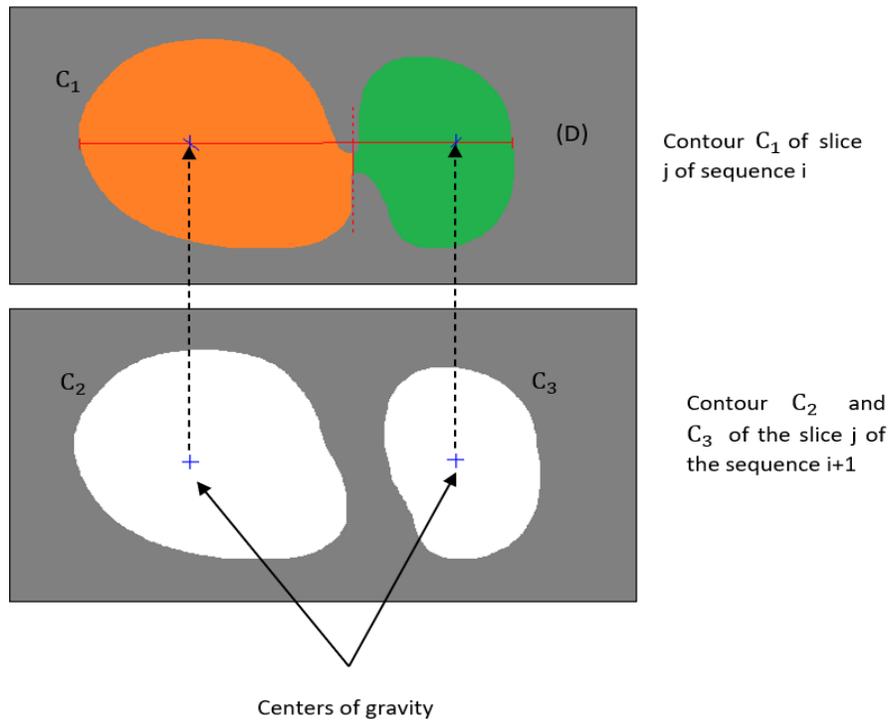


Fig. 16. Create a Correspondence between the Contour C_1 of the Cut j of the Sequence i and the Contours C_2 and C_3 of the Same Cut of the Sequence $i+1$.

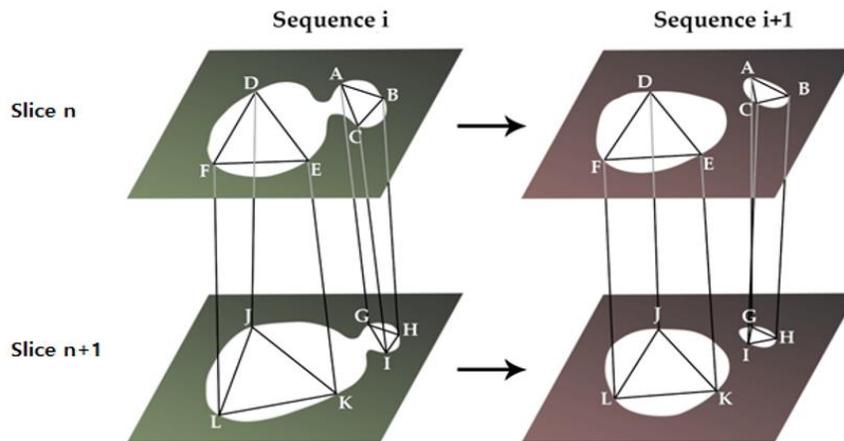


Fig. 17. Construction of the Sub-mesh which will be the Sub-mesh used from Time $i+1$.

After this operation, we can easily see that we have obtained a complete mesh associated to all the sequences (see Fig. 17). Based on such computed mesh, we can determine for each vertex at time T_i its corresponding in the next time T_{i+1} . As it can be deduced, this procedure allows the possibility to interpolate the meshes between any consecutive times.

IV. RESULTS AND DISCUSSION

The implementation of the proposed method of 4D mesh reconstruction from 2D medical images is based on the following steps:

- Detection and extraction of contours from a 2D medical image sequence.

- Construction of the basic 3D mesh of the T_0 sequence and of all the sequences that do not have some correspondence with the previous sequences.
- Extraction of displacement vectors for any sequence from the 2D images based on the corresponding 3D mesh.
- Interpolation of the 3D mesh between sequences based on the interpolation of the displacement vectors to have a smooth visualization.
- Construction of the high resolution 4D mesh from the obtained base mesh and the displacement vectors.

The first numerical result concerns the testing of our 4D reconstruction method in terms of the quality of the obtained 4D mesh. Effectively, the construction of the 4D mesh used by some classical methods is often based on the Marching Cubes algorithm combined with the interpolation method. It is worth mentioning that this approach causes ambiguity cases in some configurations [19]. In other negative situation, one may be confronted with complex areas causing an unexpected deformation during the mesh reconstruction [19]. In contrast, our new method leads to a low-resolution mesh based on contour extraction that makes it possible to create a 3D mesh without any ambiguity, matching exactly the actual shape of the anatomy. It is also noticed that the computed 4D mesh incorporate easily all the gathered anatomy information. Indeed, this shows that our method can overcome the staircase problem obtained with the classical method (see Fig. 18) (Marching Cubes with very small elementary cubes).

The second testified numerical result is related to the low storage level. For this purpose, we have compared the amount of information stored using the new adopted structure with the structure used in the conventional storage method (see Table I). As a result, by simplifying the calculations, we have applied both methods to a medical data volume of 512 x 512 x 38. This gives a size of the high-resolution mesh with the conventional method of about 2.48MB, which must be

multiplied by m (the number of given sequences). This represents a volume of about 50MB, while if we store the base mesh with the displacement vectors, we reach only 150KB. This is then a good storage reduction that is very important in practice.

The last important numerical result achieved by our algorithm is witnessed by the rendering speed. In the related numerical test, we have considered a set of 10 medical image sequences (T_0 to T_9), and in each sequence we have 38 slices. As a result, the table below shows the necessary rendering execution time for both methods (see Table II).

It is noticed that in the classical method we sent the high-resolution mesh of each time T_i directly to the GPU, while in our method we only sent to the GPU a set of basic sub-meshes with the displacement maps at each time T_n .

In another additional test, the same display is considered in both methods, by disabling in our method the interpolation between T_n time periods. The obtained comparison shows that the rendering time with the classical method is about 120 MS (8 frames per second) which is very slow about our method that achieved a rendering time equal to 19 MS only (52 frames per second). Roughly speaking, the obtained enhancement in speeding up the rendering is at least 6 times more.

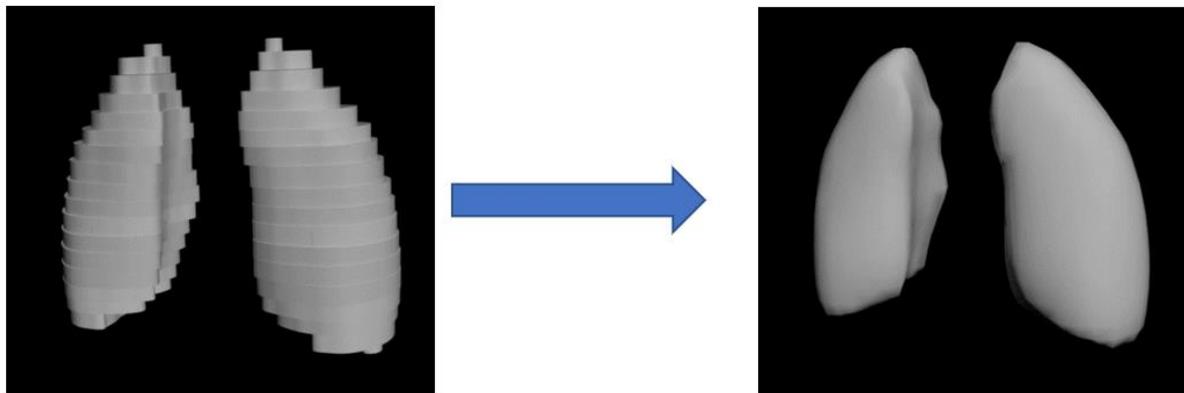


Fig. 18. Our Method Allows to Overcome the Staircase Problem Obtained with the Classical Method.

TABLE I. COMPARISON BETWEEN THE AMOUNT OF INFORMATION STORED USING THE NEW STRUCTURE AND THE CONVENTIONAL STORAGE METHOD

	Basic mesh (bytes)	Number of sequences	Displacement vectors	Total (bytes):
The original method (MC)	2,480,892	20	0	=49,617,840
The new method	126,148 (T_0) 1,009 (T_5) 1,129 (T_{12}) 1,276 (T_{16})	The other sequences are based on these 4 basics meshes already sent	500*20 vectors *2 char (x,y) *1 octet	=149,562

TABLE II. COMPARISON BETWEEN THE RENDERING SPEED USING THE NEW METHOD AND THE CONVENTIONAL METHOD

	Data sent to the GPU	Size of the sent data	Average rendering time	Number of images displayed per second
The original method (MC)	High resolution meshes 10 of size 50 MB	50 Mo	120 ms	8,33
The new method	Basic sub-mesh set and 10 displacement cards	150 Ko	19 ms	52,63

V. CONCLUSION

In this paper, we have provided an enhanced algorithm that allows the reconstruction of a 4D mesh from a set of medical image sequences, using a base mesh and displacement vectors for each sequence. The use of displacement vectors to add extra details to the basic mesh has been shown to reduce the amount of information stored in the final 4D mesh. Another established fact is that the use of the tessellation unit of the graphics card allows speeding up the rendering time; since, the proposed method create the 4D mesh, based on the 3D reconstruction method that uses the GPU tessellation unit in the rendering. This modification aims to eliminate the cases of ambiguity, especially, in certain types of objects that cannot be treated easily with the reported conventional Marching Cubes method. This is also helpful for obtaining a low-resolution mesh that can be directly used in the rendering without going through a correction step.

Another particularity of the proposed method is that one can extract the contour of the anatomy to be reconstructed from the 2D images, and then build a basic mesh within a low computational time and with a reduced storage memory. By doing so, this basic mesh construction step is done for each sequence that does not have the same contours as the previous sequence, which leads to a very small number of basic meshes. Moreover, we generate displacement vectors for each sequence by discretizing the contours according to the desired level of details.

As a result, our framework has been proven to be efficiently suitable and adaptable for the following main technical issues:

- The construction of a low-resolution base mesh using the contour extraction and matching method overcomes the staircase problem as well as the ambiguity problems and generates a mesh that exactly matches the real anatomy.
- The computation of basic meshes and displacement vectors has a very low storage burden.
- The transmission of the information to the GPU allows accelerating the rendering time.

Given the important advantages of this 4D reconstruction method, it will be very useful to apply it to the vital organ that is the heart. Thus, we intend to do detailed research and analysis to treat all its particularities.

REFERENCES

- [1] P. J. Barnes, J. Baker and L. E. Donnelly, "Cellular senescence as a mechanism and target in chronic lung diseases," *Am. J. Respir. Crit. Care Med*, vol. 200, no. 5, p. 556–564, 2019.
- [2] E. G. M. Kanaga, J. Anitha and D. S. Juliet, "4D medical image analysis: a systematic study on applications, challenges, and future research directions," *Advanced Machine Vision Paradigms for Medical Image Analysis*, pp. 97-130, 2021.
- [3] S. Zhang, G. Geng and J. Zhao, "Fast parallel image reconstruction for cone-beam FDK algorithm," *Concurrency and Computation: Practice and Experience*, vol. 31, no. 10, p. e4697, 2019.
- [4] G. McKinnon and R. Bates, "Towards imaging the beating heart usefully with a conventional CT scanner," *IEEE Trans Biomed Eng*, vol. 28, pp. 123-127, 1981.
- [5] C. Mory, *Cardiac c-arm computed tomography*, Université Claude Bernard Lyon, 2014.
- [6] S. Capostagno, A. Sisniega, J. W. Stayman, T. Ehtiati, C. R. Weiss and J. H. Siewerdsen, "Deformable motion compensation for interventional cone-beam CT," *Physics in Medicine & Biology*, vol. 66, no. 5, p. 055010, 2021.
- [7] S. Rit, J. W. Wolthaus, M. van Herk and J. J. Sonke, "On-the-fly motion-compensated cone-beam CT using an a priori model of the respiratory motion," *Medical physics*, vol. 36, no. 6Part1, pp. 2283-2296, 2009.
- [8] S. Oh and S. Kim, "Deformable image registration in radiation therapy," *Radiation oncology journal*, vol. 35, no. 2, p. 101, 2017.
- [9] G. Haskins, U. Kruger and P. Yan, "Deep learning in medical image registration: a survey," *Machine Vision and Applications*, vol. 31, no. 1, pp. 1-18, 2020.
- [10] M. J. Riblett, G. E. Christensen, E. Weiss and G. D. Hugo, "Data-driven respiratory motion compensation for four-dimensional cone-beam computed tomography (4D-CBCT) using groupwise deformable registration," *Medical physics*, vol. 45, no. 10, pp. 4471-4482, 2018.
- [11] O. Laligant and F. Truchetet, "A nonlinear derivative scheme applied to edge detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 2, pp. 242-257, 2008.
- [12] J. Canny, "A computational approach to edge detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 8, no. 6, pp. 679-698, November 1986.
- [13] S. Yi, D. Labate, G. R. Easley and H. Krim, "A shearlet approach to edge analysis and detection," *IEEE Transactions on Image Processing*, vol. 18, no. 5, pp. 929-941, 2009.
- [14] W. Gao, D. Fu, Y. Li and Y. Song, "An Image Edge Detection Method Based on Wavelet Transform Modulus Maximum," *Acta Microscopica*, vol. 28, no. 6, 2019.
- [15] J. Mairal, M. Leordeanu, F. Bach, M. Hebert and J. Ponce, "Discriminative sparse image models for class specific edge detection and image interpretation," in *European conference on Computer Vision*, Berlin, Heidelberg, 2008.
- [16] L. Miara, S. B. El Mdeghri and M. O. C. Malki, "Enhanced algorithm for reconstruction of three-Dimensional mesh from medical images using tessellation of recent graphics cards," *International Journal of Advanced Computer Science and Applications*, vol. 11, no. 12, 2020.
- [17] D. Meyers, S. Skinner and A. K. Sloan, "Surfaces from contours," *ACM Transactions on Graphics*, vol. 11, no. 3, pp. 228-258, 1992.
- [18] B. E. M. Said, C. M. M. Ouçamah and K. & Abdelhak, "Improvement of the generation of displacement vectors in the reconstruction of a 3D mesh for medical images," *International Journal of Medical Engineering and Informatics*, vol. 9, no. 1, pp. 20-29, 2017.
- [19] S. B. E. Mdeghri, M. M. O. Cherkaoui, A. Kaddari and A. Elloub, "Reconstruction of a 3D mesh with displacement vectors for medical images," *International Journal of Medical Engineering and Informatics*, vol. 7, no. 3, pp. 209-221, 2015.

An Adaptive Approach for Preserving Privacy in Context Aware Applications for Smartphones in Cloud Computing Platform

H. Manoj T. Gadiyar^{1*}, Thyagaraju G. S², R.H. Goudar³

Assistant Professor, Department of Computer Science & Engineering

Sri Dharmasthala Manjunatheshwara Institute of Technology, Ujire – 574240, Karnataka, India¹

Visvesvaraya Technological University, Belagavi, Karnataka, India¹

Sri Dharmasthala Manjunatheshwara Institute of Technology, Ujire – 574240, Karnataka, India²

Visvesvaraya Technological University, Belagavi, Karnataka, India³

Abstract—With the widespread use of mobile phones and smartphone applications, protecting one’s privacy has become a major concern. Because active defensive strategies and temporal connections between situations relevant to users are not taken into account, present privacy preservation systems for cell phones are often ineffective. This work defines secrecy maintenance issues similar to optimizing tasks, thereby verifying their accuracy and optimization capabilities through a hypothetical study. Many optimal issues arise while preserving one’s privacy and these optimal issues are to be addressed as linear programming issues. By addressing linear programming issues, an effective context-aware privacy-preserving algorithm (CAPP) was created that uses an active defence strategy to determine how to release a user’s current context to enhance the quality of service (QoS) regarding context-aware applications while maintaining secrecy. CAPP outperforms other standard methodologies in lengthy simulations of actual data. Additionally, the minimax learning algorithm (MLA) optimizes the policy users and improves the satisfaction threshold of the context-aware applications. Moreover, a cloud-based approach is introduced in the work to protect the user’s privacy from third parties. The obtained performance measures are compared with existing approaches in terms of privacy policy breaches, context sensitivity, satisfaction threshold, adversary power, and convergence speed for online and offline attacks.

Keywords—Context-aware; privacy; active defence; privacy protection and mobile phones

I. INTRODUCTION

Mobile phones are used extensively, and apps are commonly produced for smartphones. “Context-aware applications specifically help users by providing contextually relevant tailored services [1], [2]. Context-aware applications may use sensors (e.g., GPS) to determine their owner’s location and state. These sensory data may be used to determine a user’s context or condition. For example, a user’s position may be relayed via GPS, their movement assessed by accelerometers, and their voice and scene captured by cameras and microphones. Context-aware applications may use the inferred context to provide context-aware tailored services [3]. Health Monitor can track daily activity and intelligently mutes the phone without the help of the user. Context-aware applications improve people’s lives and convenience yet also

compromise privacy. Some untrusted aware applications may be highly prone to leakage of user’s context privacy to the adversary. These adversaries may sell the user privacy for commercial purposes, resulting in a reduction of QoS of the context applications. In reality, most users would not object to allowing context-aware applications to access associated sensory data because of its convenience; thus, avoiding the danger of context-privacy exposure while delivering context-related services is becoming more important [4].

Constantly changing human situations and actions in everyday life make it difficult to maintain context-privacy for mobile phones [5]. As a matter of fact, a Markov chain may represent temporal relationships across human settings. By introducing temporal correlation among present contexts can estimate the past and future usage of context applications more efficiently. Because the naïve method ignores temporal connections between user contexts, it fails to secure critical user contexts. With MaskIt [6], sensitive and non-sensitive contexts may be silenced to reduce temporal connections between them. MaskIt’s ability to hide additional contexts reduces the QoS given by context aware smartphone applications. This approach uses passive defence, which inevitably reveals some information to opponents determining whether the hidden circumstances are sensitive or not is irrelevant to an opponent. Recent proposals include aggressive defence programmes. To reduce temporal correlations across contexts, the deception approach is introduced in each context called FakeMask. FakeMask releases scenarios that are not genuine but have significance (i.e. the user has a likelihood of being in that context at that moment) [7]. An increased number of actual contexts lead to greater service quality for consumers with such a deception strategy. As a result, it is not suitable for mobile devices [8]. In recent years, many researchers have been fascinated with cloud computing [9, 10] due to the efficient outcome for protecting privacy from the third party.

A. Motivation

With the rapid advances in technology, mobile phones play an important role in users’ daily lives. However, privacy leaks are one of the most common problems in smartphones. This research mainly focuses on preserving the privacy of

*Corresponding Author.

smartphone users in context-aware applications using cloud computing. The cloud computing platform can build a firewall between the user and the adversary. Few types of research into optimal policy users have been conducted, but raising the satisfaction threshold remains a challenging approach. However, these approaches do not protect the entire privacy of the user. Mainly location and environment privacy is leaked from the smartphones for commercial purpose. Preventing an adversary's attack requires an efficient approach to better protect user privacy. Although some types of research are taking place in this field, there is also high leakage of users' privacy. These major disadvantages motivate us to develop an efficient approach to protect user privacy from adversary attacks.

A lightweight privacy preservation approach is introduced to develop temporal correlation among every user context to protect one's privacy more effectively. Furthermore, the mobile phone context privacy problem is formulated as an optimization problem and then proves its validity. It is followed by a near-optimum problem (linear programming) to speed up the execution time. By addressing the linear programming issue, an efficient context-aware privacy preserving algorithm (CAPP) is constructed that can select how to release a user's current context to optimize the QoS of context-aware applications with privacy preservation. Extensive simulations are performed to analyze the method performance, and the simulation results show that the proposed algorithm is effective and efficient. This proposed work undergoes the following major contributions to show the better QoS of the developed model:

- This research work mainly focuses on developing a novel approach for preventing leakage of data in context-aware applications for smart phones using cloud computing.
- An effective context-aware privacy-preserving (CAPP) algorithm is introduced to address the linear programming issue.
- Minimax learning algorithm (MLA) is emphasized to optimize the policy users and improve the satisfaction threshold.
- To preserve the user's privacy, the cloud computing approach is evaluated. It mainly creates a firewall between the adversary and the user.
- The evaluation of the proposed work based on privacy policy breaches, context sensitivity, satisfaction threshold, adversary power, and convergence speed for both online and offline attacks are investigated and compared with traditional approaches.

The following sections are organized as follows: Section 2 presents the literature survey related to the developed model; Section 3 describes the proposed method; Section 4 provides results and discussion; Section 5 provides the conclusion of our proposed method.

II. LITERATURE REVIEW

A. Some of the Recently Published Papers are Surveyed below

Wang et al. [11] studied the context-aware implicit authentication of smart phone users based on multi-sensor behaviour. In this method, multi-sensor data like accelerometer, gyroscope, magnetometer, time stamp, pressure and touch size were initially encapsulated to determine one's behaviour. Here, gesture and touch features were drawn from sensed data using a statistical and distance calibration approach. The features are fused using the weighted sum fusion rule, and a one-class support vector machine (SVM) was used to classify the outcome. For experimentation, 1000 sensed data from 80 participants were considered. The overall equal error rate (EER) attained was about 0.0071% in the experimental scenario. However, this method was highly suffered due to leakage of the privacy to the third party.

Alawadhi et al. [12] performed the method toward privacy protection in context aware environment. In this method, the decision making process was introduced to monitor privacy behaviour and personal data usage. This method undergoes three stages: service classifier, privacy preference manager, and privacy controller. The privacy preference module places the privacy preference and analyses the user's data usage. The next module uses service providers under trusted, untrusted and under investigation. The privacy controller detects the usage of data based on service providers. The overall false positive rate (FPR) attained was about 1.5% in the experimental scenario. However, this method suffered due to high optimization problems.

Wan et al. [13] investigated privacy preserving blockchain enabled federated learning in 5G beyond networks. This method introduced machine learning (ML) technique to keep the data efficiently. To prevent raw data sharing, a federated learning-based privacy-preserving ML has been proposed. In addition, the Wasserstein Generative Adversarial Network (WGAN) with Differential Privacy (DP) was introduced to prevent unwanted malicious attacks by interpreting the context in FL. The WGAN approach has been proposed to construct the controllable random noise that meets the DP requirements. WGAN with DP helps to achieve the trade-off between privacy and data utility. In the experimental scenario, time delay, accuracy and efficiency were calculated. However, this method preserves data for a single user and cannot be used to preserve data for multiple users.

Ghosh et al. [14] developed the context-aware security scheme to preserve the data in an IoT-enabled society. In this method, an encryption policy attribute-based encryption scheme (CP-ABE) was introduced to efficiently preserve the user's context. In addition, a context-aware attribute learning scheme (CASE) has been proposed to learn the user's contextual information and reduce the size of the encryption data by learning the attributes. The CP-ABE technique manually enforces user data by leveraging edge intelligence in

context-aware applications. In the test scenario, the delay was reduced to 33% with a packet loss of 36%. However, this method suffered from the lack of preserving the users' environmental context information.

Sylla et al. [15] investigated secure and trustworthy context management for context aware security and privacy in the IoT (SETUCOM). In this method, device trust management (DTM) was introduced based on context aware and privacy as a service (CASPaS). The context information was secured using a Bayesian network and fuzzy logic, and it was considered the lightweight hybrid system. The elliptic curve integrated encryption scheme (ECIES) algorithm generated the security. Advanced encryption standard (AES) was evaluated for context information encryption. In the experimental scenario, the overall time taken to protect the information is about 1200ms. However, this method suffered due to insecurity of the user's privacy and it's highly occurs optimization problems.

Meng et al. [16] had defined the privacy-preserving and sparsity aware location based prediction method for collaborative recommender systems. In this method, a location-based collaborative recommendation algorithm was introduced to achieve the compromise between prediction accuracy and privacy protection. A random jamming approach has been proposed to preserve data users' QoS. In addition, the regional aggregation approach was demonstrated to preserve the location of users. Furthermore, a location-based tensor factorization approach was presented to establish a relationship between services and location to develop location-based predictions. In the test scenario, the overall accuracy achieved was about 92%. However, this method suffered from poor QoS quality and takes more time to preserve the user's location.

B. Problem Formulation

Recently, many advanced techniques have been introduced to prevent user privacy leakage in contextual applications. When used context-aware, third-party intrusion is considered unusual. This happens mainly due to the lack of QoS in the existing approaches. In general, both online and offline attacks occur in context-aware applications. However, these attacks are due to attackers for commercial purposes. The literature review mentioned above greatly affects the previous approaches due to major disadvantages; some of the common disadvantages are high leakage, privacy loss and slow process, etc. Some of the existing papers have large gaps and are manipulated in this section. In [11], the author studied the context-aware implicit authentication of smart phone users based on multi-sensor behaviour to the privacy of the user's context. However, this method was highly prone to leakage of users' privacy and lack of firewall. In [12], the author performed the method toward the privacy protection in context aware environment to preserve the environmental context of the user. However, this method suffered due to optimal issues that resulted in leakage of user privacy to the adversary. In [13] the author examined privacy preservation through blockchain-enabled federated learning in 5G beyond networks. In [14] the context-aware security scheme to preserve the data in an IoT-enabled society. However, this method was highly suffered due to user's location privacy

leakage. However, this method suffered due to programming complexity to preserve one's context privacy. In [15], the author proposed secure and trustworthy context management for context aware security and privacy in the IoT (SETUCOM) to improve the QoS of the context applications. However, this method suffered due to the high complexity process and the lack of preserving location privacy more effectively. In [16], privacy-preserving and sparsity aware location based prediction method for collaborative recommender systems has been proposed by the author. However, this method is only useful for getting the user's location.

Few researchers were undertaken to protect users' privacy based on cloud computing. The aforementioned related paper is efficient and shows a better outcome in terms of privacy preservation, but there also arises some leakage due to less improvement in the applied strategy. An effective novel approach needs to be introduced to preserve privacy to overcome this issue. This proposed method gives a clear solution for data protection with better accuracy.

III. PROPOSED METHOD

Context aware is the computation of the current situation and information about the environment, places, things that anticipate urgent needs, situate awareness, usable contents and experiences. In this work, a novel approach is developed to prevent an attack from an adversary concerning user's privacy. Initially, CAPP algorithm is introduced to enhance the quality of service (QoS) regarding context-aware applications. Then, the MLA algorithm is emphasized to equalize the optimal policy and improve the satisfaction threshold of the proposed work. The proposed work introduces cloud computing to encrypt users' privacy from a third party. Fig. 1 illustrates the framework of the proposed model.

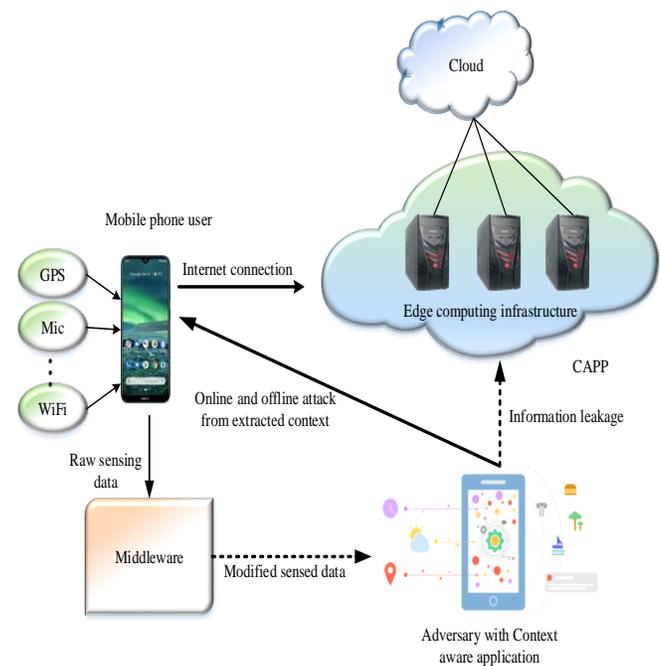


Fig. 1. Framework of the Proposed Method.

Cloud computing is the collection of network servers coordinated with the aid of the internet. The cloud uses firewalls as privacy protection around assets to prevent intruding of third parties. To carry out this evaluation, real smartphones with traces of 94 users to find out the convergence speed of the algorithm. This paper mainly focuses on online and offline attacks due to continuous changing times and user variability.

A. Privacy Problems in Context Aware Privacy Preservation

The private contexts are said to be the context subsets in which leakage is considered the major drawback for the smartphone user. In order to prevent the leakage of privacy, the user must control the emitted information using middleware privacy preservation. Many existing approaches are introduced to overcome the leakage of the user's privacy. Privacy-preserving middleware is used to access the context-aware middleware for the users, but this middleware does not require any permission to access the user's data. Usually, the released data with granularity leaks the privacy of the user. Hence the accuracy of the context recognition is also reduced. The context-aware apps are mainly used for commercial purposes and are considered the adversary. The adversary is a third party intrusion, mainly focusing on reducing the user's utility by adding multiple attacks. The attacks mainly undergo two stages offline and online attacks.

The third party gets the user's personal information such as behavioral contexts, GPS location information, etc. These third parties sell the information for commercial purposes, and users are unaware that the attack leads to a privacy breach. In online attacks, the third party collects the sensed data from the user and understands the user's behavior based on the collected data. According to the behavior, the third party forces the user or makes the user indulge in blackmail or leads to violence.

1) *State transition for online attacks:* For the online attack, the adversary's strategy is unknown to the user, and the user blindly believes the adversary's strategy from the existing attacks. There are many reasons for the dependence on previous attacks. The third party collects the last context based on the previous approach. Then the present information is coordinated with the past information based on the proposed algorithm. Hence the user should encapsulate the adversary's attacks and which information has been attacked. The attacked time is denoted by n , which clearly shows the time the data gets leaked. The attack of the previous information is indicated as Wr^n , the value of Wr^n is 1 or 0. If the adversary successfully collects the information is denoted as D_{n-1} . The state transitions for the online attacks at the time n is given by, $R^n = \{Wr^n, D_{n-1}\}$.

2) *State transition for offline attacks:* In an offline attack, the adversary gets the user's personal data such as personal

behavior contexts, environmental context, GPS location etc. The adversaries sell one's personal information for money and lead to a privacy breach, and it is unknown to the user. The time n of the user's action is given by, $a_h^n = \{a_{h,1}^n, \dots, a_{h,E}^n\}$, the granularity of the sensor's data is denoted as $a_{h,E}^n \in [0,1], \forall E=1, \dots, E$. Here, E denotes the complete sensors for the purpose of recognition. The recognition of the context in terms of accuracy with the limit ranges from t ($0 \leq t \leq 1$) is given by,

$$t = \sum_{E=1}^E E_e a_{h,E}^n \quad (1)$$

Here, $\{E_e : \forall E\}$ denotes the weight of the context sensitivity based on context recognition accuracy.

The adversary's attacking capability needs to choose the correct subset of regretting the sensed data. A formula gives the time n with the adversary actions as,

$$a_b^n = \{a_{a,1}^n, a_{a,E}^n\} \quad (2)$$

Here, $a_{b,e}^n$ denotes the E th sensor of the retrieved data. The adversary actions with limited power adversary given mathematically as,

$$\sum_E a_{b,j}^n \leq L, 0 \leq a_{a,j}^n \leq 1, \forall E \quad (3)$$

Here, L denotes the adversary power limitations. Based on the limit $L \leq E$, the third party can capture the sensed data. Hence this adversary is said to be an unlimited power adversary.

The behavior of the adversary in online attacking the user can be determined in a probabilistic manner, and it is given by,

$$Pb[R^n | a_h^n, a_a^n] = Pb[Wr^{n+1} | Wr^n, a_h^n, a_a^n] Pb[D^{n+1} | D^n] = Pb[Wr^{n+1} | a_h^n, a_a^n] Pb[D^{n+1} | D^n] \quad (4)$$

The offline attack in case of adversary based on the probabilistic manner is given by,

$$Pb[R^{n+1} | a_h^n, a_a^n] = Pb[D^{n+1} | D^n] \quad (5)$$

B. Proposed Context Aware Privacy Preservation Algorithm

The optimization problem gets converted into a linear programming problem to improve the convergence speed. To overcome the linear programming problem, the CAPP algorithm is introduced. It generates the active policy users and increases the service quality for context aware applications with user privacy.

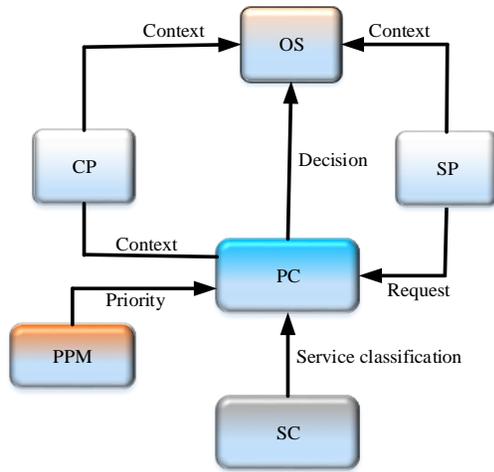


Fig. 2. Architecture of Privacy Aware.

Fig. 2 shows the architecture for privacy aware. The service provider gives a request to the PC about the priority. The PC does not request permission from the user directly and sends the requests through the PC module. The module provides the decision and notifies it to the OS that communicates with the SP consecutively. The user's privacy is stored in the privacy preference manager (PPM). It enables the user to set the privacy and sensitivity to context aware apps.

- Dividable.
- Not dividable.
- To be established.
- Cannot be established.

Dividable means the user feels comfortable for giving one's personal information. Not dividable denotes that the user feels insecure in sharing one's information. The data to be established depicts that when the user uses the application source for the first time; he sets the location permission to be dividable or cannot be dividable to the application. It cannot be established indicates that the user is unable to express to set the context priority that is sharable or not. The SC module describes the category of the service provider based on the context request and how it is divided among the SP. SP undergoes 3 stages: hopeful, hopeless, and examining. The hopeful SP is the one who asks only necessary information from the user. The hopeless SP is the one who asks for unnecessary that are not required for the adversary. The adversary is tested for hopeful or hopeless SP in the investigation category.

The PC module senses the user when the hopeless SP attacks the personal data. The deciding operation is done by three operation:

- Allow – allow access to context request.
- Deny – prevent access from context requests.
- Approximation- allow access to approximate the value based on the context request.

Algorithm For Context Aware Privacy Preservation

```

Step 1: Initialize service provider (SC)
Step 2: Allow request for PC to preference
Step 3: if, users data is sharable
  Allow PC to get service provider from SC
Step 4: else, users data not sharable
  Send decision to OS
  End
Step 5: if, SP is trusted
  Allow permission to access users data
  Send decision to OS.
  End
Step 6: else, SP is not trusted
  Access deny
  Send decision to OS
  End
Step 7: do, investigation and approximate range of SP
  Send decision to OS
  End
Step 8: else,
  Allow request for PC to preference
Step 9: if, SP is trusted
  Suggest to user to access the data
  Gets users decision from PC
  Send decision to OS
  End
Step 10: else,
  Suggest not to user to access the data
  Gets users decision from PC
  Send decision to OS
  End
Step 11: else,
  Suggest to user to set the approximate range
  Gets users decision from PC
  Send decision to OS
  End
  
```

The algorithm selects how to reveal a user's context while protecting their privacy. Even if an opponent knew the Markov model with associated probability matrices of emissions, they couldn't determine the original context from CAPP's output context sequence. Because privacy ensures a condition with which an adversary can never know the user when within the sensitive environment.

C. Mini Max Learning Algorithm

The minimax algorithm is a step by step process that aids the computer in working intelligently instead of not learning automatically. It mainly helps to improve the satisfaction threshold for the proposed approach.

The pair of the optimal policy is given by, $\phi^* = \{\phi^*_h, \phi^*_a\}$ for the game based context privacy is achieved by overcoming the convergence problem.

$$\tilde{U}^{n+1}(Wr) = (1 - \beta^{n+1})\tilde{U}^n(Wr) + \beta^{n+1}G_i[s_h(r, a^n_h, a^n_a) + \lambda\tilde{U}^n(Wr')] \quad (6)$$

Based on the eqn (1), using the learning approach, the $\tilde{U}^{\phi^*}_h(Wr)$ is obtained, and this can be obtained using upgraded rule based on the Q-learning approach.

$$\phi^* = \arg \max_{\phi_h} \min_{\phi_a} \{s_h(r, a^{\phi^*}) + \lambda \sum_{Wr'} (Tr[Wr' / a^{\phi^*}] \tilde{U}^{\phi^*}(Wr'))\} \quad (7)$$

The algorithm of minimax learning is shown below:

Mini-max learning algorithm

Input: the stochastic game for context aware privacy given by, Ψ

Output:

// (i) Start

1. $n \leftarrow 0, W_r^n = 0;$

2. $\tilde{U}^n(W_r=0) \leftarrow 1, \tilde{U}^n(W_r=1) \leftarrow -1;$

3. Start the pair for policy ϕ^n : distributed based on uniform

$$a^{n_h,j} = \frac{1}{P}, a^{n_a,j} = \frac{L}{P}, \forall j;$$

// (ii) Recursion

4. Iterate

5. Choose the action pair $\{a^n_h, a^n_a\}$ according to ϕ^* ;

6. Upgrade W_r^{n+1} after each user's taken into consideration as

$$\{a^n_h, a^n_a\}$$

7. Upgrade $\tilde{U}^{n+1}(W_r)$ convergence state notation as, based on equation (2).

8. Upgrade the optimized policy as ϕ^{n+1} based on the equation (1) with upgraded stated notations;

9. $n \leftarrow n + 1;$

10. Until equalize

The learning rate is indicated as, $\beta^n \in (0,1)$ for the convergence of learning algorithm, must degrade the high time operation. Set $\beta^n = \frac{1}{n} \tilde{U}^{n+1}(W_r)$ that is used as an approximate value and updated continuously until it equalizes.

Algorithm 1 evaluates the learning algorithm in an equivalent state, denoted as, $\tilde{U}_h^{\phi^*}(W_r)$. Initially, set the equalization state values as 1 and make the uniform distribution among the players of each policy. After that, the equivalent state values are continuously repeated based on equations (1) and (2). This repetition helps to occur optimal policy among the policy pair.

IV. RESULTS AND DISCUSSION

The context-aware privacy-preserving algorithm (dubbed CAPP) was developed and compared with currently present algorithms of privacies such as EfficientFake [8] and MaskSensitive, MaskIt (using the hybrid check) [7]. Basic method's MaskSensitive, which hides or suppresses all sensitive circumstances during the release of a non-sensitive one. The entire simulation was done using MATLAB. Initially, an effective context-aware privacy-preserving (CAPP) algorithm is introduced to address the linear programming issue. Then, minimax learning algorithm (MLA) is emphasized to optimize the policy users and to improve the satisfaction threshold. The performance measures such as privacy policy breaches, context sensitivity, satisfaction threshold, adversary power, and convergence speed for both online and offline attacks are investigated.

A. Analysis of Performance Metrics for the Proposed Model

To analyze the performance of the proposed method, a Markov chain is given to each user to train and assess protect the privacy context for every user. Because of the inadequacy

of previous beliefs and the probability of emission, privacy may not be ensured while gathering the user's trace. To ensure user privacy is maintained, the privacy value, which is set to 0.1, is used as the simulation parameter. It was to be noted as a higher privacy parameter, and then lower will be user privacies guarding levels with additional actual sensitive information being revealed. Selecting sensitive environments may be done in one of two ways. Unless, sensitive circumstances pertaining to every user is selected randomly, referred to as sensitive, unless otherwise noted. Alternatively, for each user, a random place is selected with the greatest probability of prior as the user's house, marking that sensitive, dubbed home as sensitive. Because the expected amount in released real context was utility about privacy-preserving technique, normalized utility is utilized as evaluation, defined as the proportion of release actual context. It's worth noting that context-aware applications give better service when their utility is greater. Identically, the amount of sensitivity contextual is splitted within the user's context sequence, which got discontinued by the user's context sequence length in evaluating privacy breaches. Three Methods: MaskIt [6], CAPP [16] and Efficient Fake [17], and everyone guarantee no violation of privacy, according to the definition. Mask because of the lack of assumption for the presence of temporary connections across the user's context, Sensitive is unlikely to be able to ensure the desired privacy.

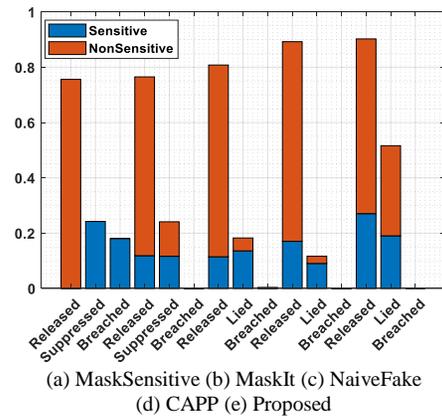


Fig. 3. Comparison for Privacy Policy Breaches (Home as Sensitive).

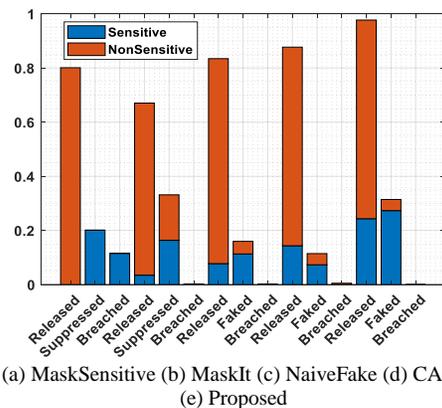


Fig. 4. Comparison for Privacy Policy Breaches (Sensitive as Random).

In the first example, comparisons are made for the privacy of CAPPs and MLA to violate the use of these alternative techniques. With certain conditions, three contexts were chosen by us randomly as sensitive about every user, whereas, in the other, each user's house is selected as sensitive. It's worth noting that a user's house has the greatest previous belief, indicating that the user must spend most of their time inside the house rather than elsewhere. In the preceding two instances, Fig. 3 and 4 shows released and repressed contexts in average proportions by different methods. All sensitive circumstances were surpassed by Mask Sensitive in entire instances, as can be seen in the images. Even though not every sensitive context was revealed within Mask Sensitive, opponents that understand the contexts of the Markov chain can predict around 40–60% sensitive contexts out of suppressed ones in both cases. The major reason is that the temporal connection across contexts gives an adversary enough information to conclude a greater post belief that can surpass correspondingly preexisting belief by privacy parameter. On the other hand, CAPP, EfficientFake, and MaskIt ensure that -privacy is maintained. Some sensitive and non-sensitive contexts are repressed and released in CAPP, EfficientFake, and MaskIt. CAPP also releases a higher percentage of genuine situations than MaskSensitive, MaskIt, or EfficientFake. MaskIt compromises less than 20% of Mask Sensitive's functionality to ensure anonymity, as seen in the numbers.

However, compared to Mask Sensitive, both EfficientFake and CAPP boost usefulness by about 20% while ensuring anonymity. The fundamental reason for this is that the new deception strategy makes it harder to antagonist in deriving posterior beliefs, allowing more genuine contexts to be released. Despite the fact that CAPP and EfficientFake were both formalized in linear programming problems, the proposed CAPP outperforms Efficient Fake techniques with both instances in terms of average utility. There are two primary reasons for this. The first difference is that in EfficientFake, the aim is to optimize emission probability solely. Still, with CAPP, the aim was to maximize the value of utility for a provided period of time. Secondly, Efficient Fake's space of resolution has shrunk significantly. The emission probability matrix' Shape in EfficientFake was reduced to a vectored representation, thereby significantly reducing the solution's precision in EfficientFake, resulting in lower utility than CAPP. On the other hand, CAPP does not shrink the solution space, allowing us to find a superior optimized solution.

The usefulness of the proposed CAPP is then compared to that of other techniques with various privacy settings ranging from 0.05 to 0.3. Separate sensitive settings are selected in the trials, just as in the previous ones: the sensitive environment for one person is their home, while the other is chosen at random. With the reduction in the privacy requirement, anticipate utility need to be rise. Fig. 5 and 6 shows that utility grows slowly as the number of people rise in both circumstances.

Moreover, the experimental analysis shows that, for similar privacy values, every solution executes in the best way within 2nd case. The context of random was designated

sensitive, compared with the first situation, when the home was designated as sensitive. Because locations of every having greatest belief of prior were picked sensitive context in the first scenario in Fig. 5, the number of sensitive contexts was greater than the 2nd situation in Fig. 6, in which sensitive context was selected at random. CAPP and Efficient Fake should release more fake contexts in the first case to give identical privacy levels. However, when related to other methods, the proposed CAPP outperforms them all due to its close approximation of the problem's optimum solution.

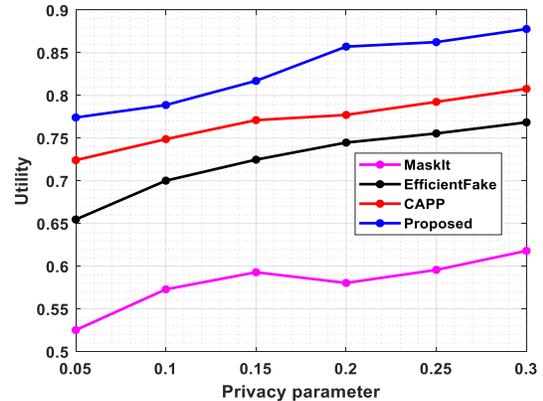


Fig. 5. Tradeoff for Privacy-utility (Home as Sensitive).

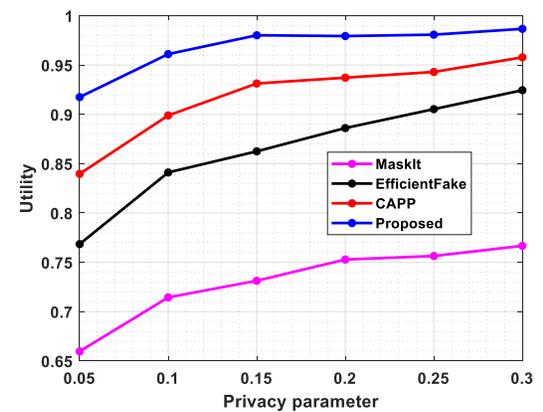


Fig. 6. Privacy-utility Tradeoff (Sensitive as Random).

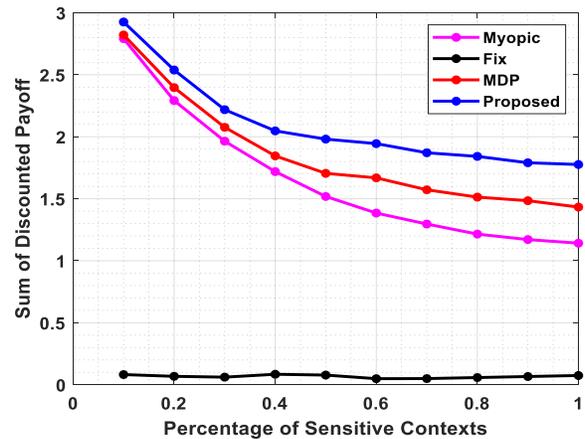


Fig. 7. Comparison of Sensitive Context and Payoff Discount.

Fig. 7 represents the comparison of sensitive context and the sum of the discounted payoff for an online attack. When the sensitive context percentage increases, the myopic strategy [18] gets highly decreased. The users of high sensitive contexts result in more privacy leakage. But this strategy shows high leakage in the case of the privacy policy. In the fixed strategy [19], the same percentage of sensitive context results in the same sum of the discounted payoff. However, this strategy shows poor outcomes because of its insecure privacy. It is due to the constant quality of service, and it controls the payoff discount in case of a fixed strategy. In the case of the MDP approach [20], it is the same as the myopic strategy.

However, this approach works more efficiently in the case of privacy preservation than the myopic strategy. But, this method shows a lie outcome based on context aware privacy preservation. The proposed model shows better results in protecting the privacy of the user. By increasing the percentage of the sensitive context to 1, the sum of the payoff discount gets very much increased to 2.9 because of using an optimal algorithm with cloud computation.

Fig. 8 compares the satisfaction threshold and sum of the discounted payoff. The satisfaction threshold is compared with the existing application based on the sum of discounted payoff. From the graph, it is shown that if sum of the discounted payoff gets lower, the satisfaction threshold gets very much lower. If the satisfaction threshold gets increased, the quality of service gets lower. The existing approach results in low-quality service by comparing the proposed model with the myopic strategy. If the service quality decreases, it is harder to better accuracy in the outcome. Hence the satisfaction threshold must be lower to prevent the loss of leakage in privacy. In the case of the MDP approach, there attains a better quality of service. However, this method shows some drawbacks in hiding the information from the third party. The proposed model shows better privacy protection as the satisfaction threshold is only 0.15.

Fig. 9 represents the different context sensitivity based on optimal police based on online policy. Here, a, b illustrated in the graph denotes the released and leaked data. With the smaller context sensitivity of 0.25, the optimal policy achieves 1. When the context sensitivity becomes higher by about 0.87, the optimal policy attains a negligible value. In the case of smaller context sensitivity, the service quality variance improves, and vice versa, the loss of privacy dominates. Suppose both a and b go down, the context sensitivity increases. Due to this, the user chooses a more optimized strategy to protect their privacy efficiently. From the graph, it is clear that if the satisfaction threshold increases, the context sensitivity decreases. Suppose the user receives only low-quality service if the threshold becomes lower. Understanding the satisfaction threshold is considered an important parameter while developing context-aware privacy protection approaches.

Fig. 10 illustrates the different adversaries with optimal policies for the offline attack. Here, L denotes the adversary power of the privacy policy. Considering, at L=1, attains poor adversary power due to high leakage of privacy of the user. In

myopic strategy, the discounted payoff with different adversary power is higher than the proposed work. For efficient usage of context aware applications, the sum of discounted payoff should be too low. From the graph, it is shown that the sum of discounted payoff gets degraded, resulting in no leakage. In existing approaches, compared with the limited adversary power, the performance of the adversary power with unlimited power gets very worse. Mainly, the adversary with unlimited power occurs more leakage, and hence the user selects another strategy to protect their privacy.

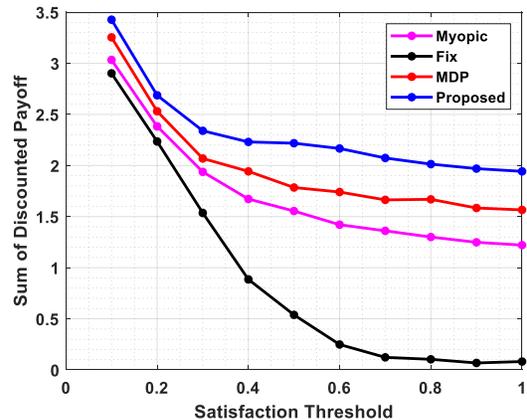


Fig. 8. Comparison of Satisfaction Threshold and Payoff Discount.

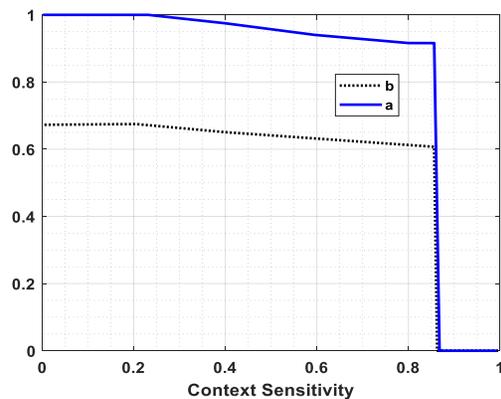


Fig. 9. Different Context Sensitivity based on Optimal Policies.

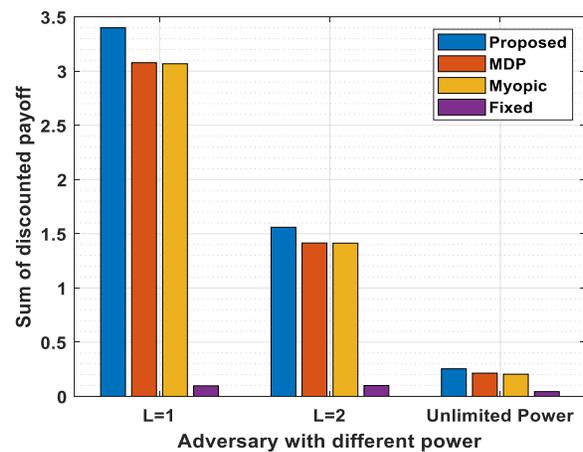


Fig. 10. Different Adversaries with Optimal Policies.

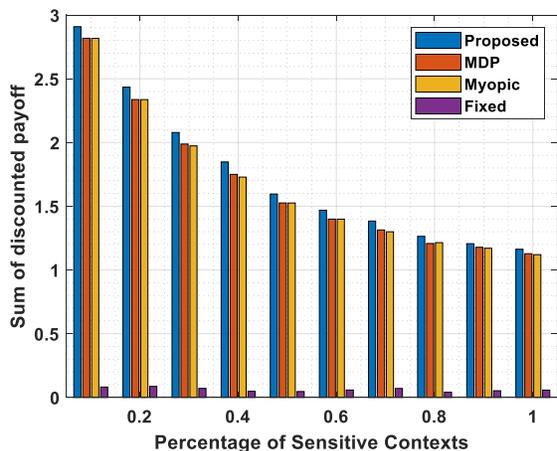


Fig. 11. Comparison of Sensitive Context and Payoff Discounts.

Fig. 11 illustrates the comparison of sensitive context and payoff discounts. As per the graph, when the sensitive context percent increases, the discounted payoff's sum gets degraded. This denotes that sensitive users have to use the more encrypted forms to preserve user privacy. When the sensitive context percent increases in myopic strategy, the discounted payoff's sum gets degraded slowly. But this approach does not show better accuracy in preserving the user's privacy under offline attacks. In fixed strategy, the sum of discounted payoff gets diminished gradually with an increase in sensitive contexts. But this approach uses high complexity in protecting one's privacy from the third party. The MDP approach is also the same as the other existing approach because of the lack of new approaches to protect the privacy of the user. The proposal shows a better outcome because it performs effectively in preserving the privacy of the user.

Fig. 12 compares the satisfaction threshold and sum of the discounted payoffs. As per the graph, when the satisfaction threshold (accuracy) increases, the sum of the discounted payoff gets decreases slowly in the case of myopic strategy. This shows that this strategy is not suitable for preserving the privacy of the user due to low service quality. In the case of fixed strategy, the sum of discounted payoff decreases significantly with increased accuracy. This approach shows a lack of service quality due to the absence of a slow encryption process. Considering the MPD approach, the sum of the discounted payoff gets decreased with an increase in accuracy but is not efficient due to a lack of preserving privacy from the third party. When the sum of the discounted payoff gets decreased, the accuracy is very high. This shows that the proposed method with cloud computing helps the user protect their privacy effectively.

Fig. 13a, 13b, 13c shows the sum of discounted payoff at $L=1$, $L=2$ and with unlimited power. As per the graph, the discounted payoff gets reduced when the adversary's power gets increased. It is due to the increase of L , and the adversary can access more data. Hence it is influenced by the adversary to attack the user successfully. Considering the releasing data with less granularity shows the lower service quality or the user should depend on the same approach to protect user's privacy. This leads to more loss in privacy of the user and less

payoff. The existing approaches like myopic, fixed and MDP approach more leakage in the privacy of user due to low satisfaction threshold. The proposed method shows good encryption in protecting the user's privacy because of high satisfaction threshold.

Fig. 14 illustrates the cumulative distribution function (CDF) of various iterations in order to learn the optimal policy of the reality mining dataset. The operating speed of the proposed work is analyzed for 220 iterations. For the MPD process, the convergence speed is analyzed with 10^5 iterations. This shows that the proposed algorithm has a higher convergence speed than the MDP process. The proposal algorithm's equivalent state value helps reduce the high dimensionality for learning the process efficiently.

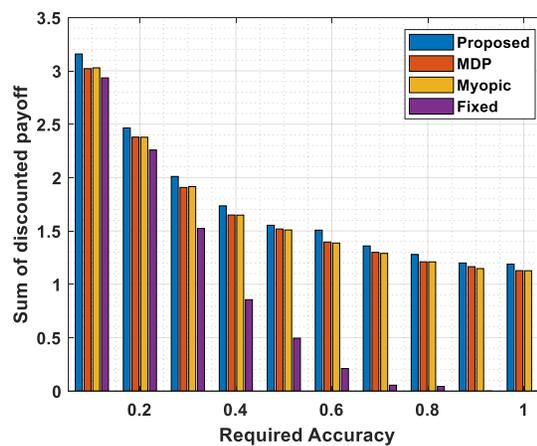
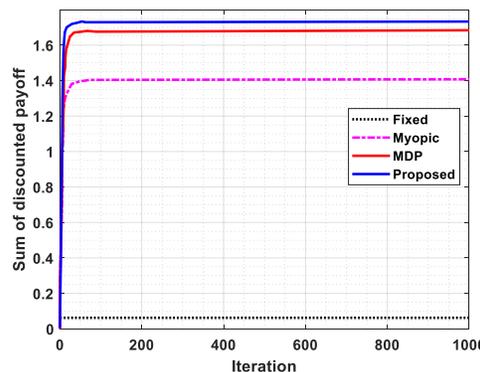
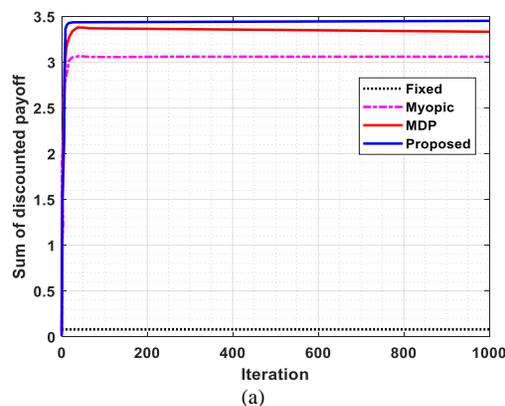


Fig. 12. Comparison of Satisfaction Threshold and Payoff Discounts.



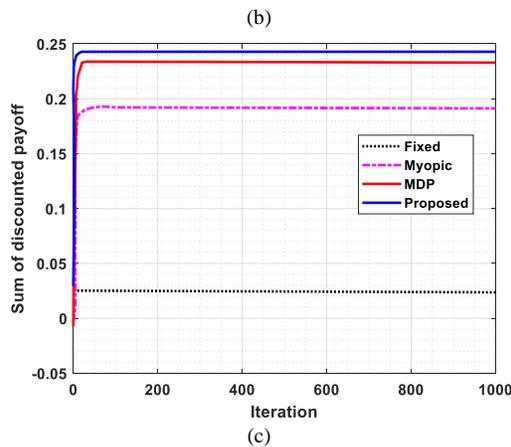


Fig. 13. Comparison of Payoff Discount and for Varying Iteration. (a) Sum of Discounted Payoff at L=1. (b) Sum of Discounted Payoff at L=2. (c) Sum of Discounted Payoff for Unlimited Power.

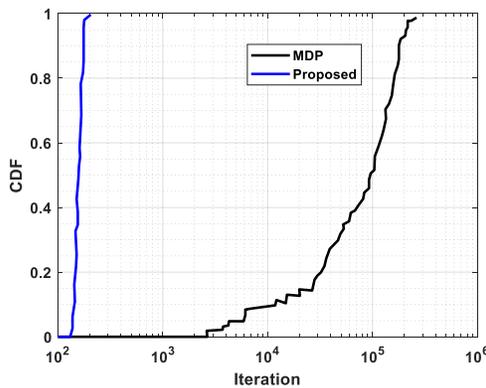


Fig. 14. Convergence Speed based on CDF.

B. Discussion

This research mainly focuses on developing a novel approach for preventing data leakage in context aware applications for smart phones using cloud computing. An effective context aware privacy preserving (CAPP) algorithm is introduced to address the linear programming issue. Minimax learning algorithm (MLA) is emphasized to optimize the policy users and improve the satisfaction threshold. To preserve the privacy of the user, the cloud computing approach is evaluated. It mainly creates a firewall between the adversary and the user. The performance of the proposed work is compared with existing approaches like Naivefake, MaskIt, EfficientFake and CAPP models. But this method suffered due to multiple drawbacks. To protect the privacy of the user is not an easy task. Many issues like the third party intrude and cracking may occur due to the lack of advancements in sensor networks.

In the context-aware implicit authentication of smartphone users based on the multi-sensor behaviour, this method was suffered due to leakage of the privacy to the third party. EER attained was about 0.0071%. In the method towards the privacy protection in context aware environment; however, this method suffered due to high optimization problems in context aware scalable authentication (CASA). However, this method suffered due to high attacks from the adversary. FPR

attained was about 1.5%. An aware access control framework for software services (PO-SAAC) was introduced in the purpose-oriented situation. However, this method suffered due to high computational complexity and increased memory size. The memory size utilized about 1600 KB based on the response time in the secure and trustworthy context management for context aware security and privacy in the IoT (SETUCOM). However, this method suffered due to insecurity of the user's privacy and it's highly occurs optimization problems. The overall time taken to protect the information is about 1200ms. The evaluation of the proposed work based on privacy policy breaches, context sensitivity, satisfaction threshold, adversary power, and convergence speed for both online and offline attacks are investigated and compared with traditional approaches. Because of its outstanding performance avoids leakage and privacy loss in smartphones because of its outstanding performance.

V. CONCLUSION

The challenge of context-aware privacy preservation for cellphones is addressed in this study. The validity and optimality is verified by theoretical analysis by formalizing the problem of contextual privacy preservation as an optimization problem. To further speed up the computation, an effective, near-optimized strategy involving the linear programming problem was developed. A context-aware privacy preserving algorithm (CAPP) was presented due to the linear programming issue being solved. The experimental analysis proves that the suggested CAPP provides much more value than existing techniques while respecting the user's δ -privacy policy via thorough experimental evaluations on actual mobility traces. A cloud-based approach is introduced in this work to protect the privacy of the user from a third party. In addition to this, a minimax learning algorithm is emphasized for improving the accuracy of the context aware application and improving the optimal policy of the users. The performance measures obtained are compared with existing approaches in terms of privacy policy breaches, context-sensitivity, satisfaction threshold, adversary power, and convergence speed for online and offline attacks.

A. Future Scope

An exciting future project would be the development of an online context released judgment system that can generate faster and most effective judgments depending just on the user's current context while maintaining anonymity. Because this research focuses on preserving privacy for a single user, a future project will provide a privacy preservation technique that considers user interactions, given that people have group mobility.

ACKNOWLEDGMENT

I sincerely thank Thyagaraju G. S, R H Goudar, for their guidance and encouragement in carrying out this research work.

REFERENCES

- [1] T. Hussain, and R. Alawadhi, "A Privacy Protection System in Context-aware Environment The Privacy Controller Module," Proceedings of the 22nd International Conference on Information Integration and Web-

- Based Applications & Services, 2020.
- [2] J. Shu, R. Zheng, and P. Hui, "Cardea: Context-aware visual privacy protection for photo taking and sharing," Proceedings of the 9th ACM Multimedia Systems Conference, 2018.
- [3] W. Ali, R. Kumar, Z. Deng, Y. Wang, and J. Shao, "A federated learning approach for privacy protection in context-aware recommender systems." The Computer Journal vol. 64, no. 7, pp. 1016-1027, 2021.
- [4] L. Gao, T.H. Luan, B. Gu, Y. Qu, and Y. Xiang, "Context-Aware Privacy Preserving in Edge Computing." In Privacy-Preserving in Edge Computing, Springer, Singapore, pp. 35-63, 2021.
- [5] Y. Zhang, J. Pan, L. Qi, and Q. He, "Privacy-preserving quality prediction for edge-based IoT services." Future Generation Computer Systems vol. 114, pp. 336-348, 2021.
- [6] J. Al-Muhtadi, K. Saleem, S. Al-Rabiaah, M. Imran, A. Gawanmeh, and J.J.P.C. Rodrigues, "A lightweight cyber security framework with context-awareness for pervasive computing environments." Sustainable Cities and Society vol. 66, pp. 102610, 2021.
- [7] V. Stephanie, M.A.P. Chamikara, I. Khalil, and M. Atiquzzaman, "Privacy-preserving location data stream clustering on mobile edge computing and cloud." Information Systems vol. 107, pp. 101728, 2022.
- [8] A. A. Ahmed, "A privacy-preserving mobile location-based advertising system for small businesses." Engineering Reports vol. 3, no. 11, pp. e12416, 2021.
- [9] T. N. Phan, et al, "A context-aware privacy-preserving solution for location-based services," 2018 International Conference on Advanced Computing and Applications (ACOMP). IEEE, 2018.
- [10] E. Ezhilarasan and M. Dinakaran, "Privacy preserving and data transpiration in multiple cloud using secure and robust data access management algorithm." Microprocessors and Microsystems vol. 82, pp. 103956, 2021.
- [11] R. Wang and D. Tao, "Context-aware implicit authentication of smartphone users based on multi-sensor behavior," IEEE Access, vol. 7, pp. 119654-119667, 2019.
- [12] R. Alawadhi and T. Hussain, "A Method Toward Privacy Protection in Context-Aware Environment," Procedia Computer Science, vol. 151, pp. 659-666, 2019.
- [13] Y. Wan, Y. Qu, L. Gao, and Y. Xiang, "Privacy-preserving blockchain-enabled federated learning for b5g-driven edge computing." Computer Networks vol. 204, pp. 108671, 2022.
- [14] T. Ghosh, A. Roy, S. Misra, and N.S. Raghuvanshi, "CASE: A context-aware security scheme for preserving data privacy in IoT-enabled society 5.0." IEEE Internet of Things Journal 2021.
- [15] T. Sylla, M.A. Chalouf, F. Krief and K. Samaké, "SETUCOM: Secure and Trustworthy Context Management for Context-Aware Security and Privacy in the Internet of Things," Security and Communication Networks, vol. 2021, 2021.
- [16] S. Meng, L. Qi, Q. Li, W. Lin, X. Xu, and S. Wan, "Privacy-preserving and sparsity-aware location-based prediction method for collaborative recommender systems." Future Generation Computer Systems vol. 96, pp. 324-335, 2019.
- [17] H. Vahdat-Nejad, S. Izadpanah and S. Ostadi-Eilaki, "Context-aware cloud-based systems: design aspects," Cluster Computing, vol. 22, no. 5, pp. 11601-11617, 2019.
- [18] X. Ding, R. Lv, X. Pang, J. Hu, Z. Wang, X. Yang, and X. Li, "Privacy-preserving task allocation for edge computing-based mobile crowdsensing." Computers & Electrical Engineering vol. 97, pp. 107528, 2022.
- [19] X. Wang, S. Garg, H. Lin, G. Kaddoum, J. Hu and M.S. Hossain, "PPCS: An Intelligent Privacy-Preserving Mobile-Edge Crowdsensing Strategy for Industrial IoT," IEEE Internet of Things Journal, vol. 8, no. 13, pp. 10288-10298, 2020.
- [20] B. D. Deebak, and A-T. Fadi, "Privacy-preserving in smart contracts using blockchain and artificial intelligence for cyber risk measurements." Journal of Information Security and Applications vol. 58, pp. 102749, 2021.

Digital Learning Tools for Security Inductions in Mining Interns: A PLS-SEM Analysis

José Julián Rodríguez-Delgado¹, Patricia López-Casaperalta²
Mario Gustavo Berrios-Espezúa³, Alejandro Marcelo Acosta-Quelopana⁴, José Sulla-Torres⁵
Escuela Profesional de Ingeniería de Minas, Universidad Católica de Santa María, Arequipa, Perú^{1, 2, 4}
Universidad Nacional de San Agustín de Arequipa, Arequipa, Perú³
Escuela Profesional de Ingeniería de Sistemas, Universidad Católica de Santa María, Arequipa, Perú⁵

Abstract—New pedagogical tools have been introduced in educational contexts in recent years. They have been shown to impact learning compared to conventional education strategies positively. Before implementing new learning tools, a study of technological acceptance is needed for its application to succeed. For this reason, the objective of this research was to measure the intention and acceptance of the use of new digital learning tools, such as mobile applications, holograms, interactive platforms, and virtual or augmented reality, through the Technology Acceptance Model 3 (TAM3) in safety on-board training inductions in a mining company. This measurement was based on the analysis of a survey carried out in Google Forms based on the Likert scale; the results were processed using the partial least squares technique in structural equation models (PLS-SEM), processed through SmartPLS 3. As a result, we got positive correlations between the instrument's variables and acceptance by the participants studied. The findings indicate that it is essential to consider the participants' opinions a priori to implementing new digital education tools for managerial decision-making. It was considering highlighting the teaching about safety in mining companies since this allows contributing to engineering education and protecting the most precious resource of any company, the human being.

Keywords—Technology acceptance model 3; PLS-SEM; SmartPLS; mining company; safety; safety inductions; safety talk; learning; education; teaching; technology; learning tools

I. INTRODUCTION

Technology is developing exponentially; it has covered fundamental aspects of human beings such as learning, focusing on eight topics: institutional environment, presence, pedagogy, technological elements, design, behaviors, affective elements, and learning outcomes [1]. It is worth mentioning that new technologies are gradually replacing the traditional approach to teaching [2] with a more social learning approach [3] as the number of publications about how the use of technology can be successful so that teachers can teach effectively, taking advantage of the opportunities of technological advances, generating an innovative learning environment based on play. Awakening students' interest in its use will better understand information [4][5]. However, research on technology-enhanced learning has been written chiefly by authors working in the educational field,

representing approximately 70% of articles in all studies collected by a bibliometric analysis carried out in 2019; on the other hand, so only 9% were from articles written by authors who are working in engineering areas [6]. There is a lack of attention to applying digital learning tools in engineering, such as the training and inductions directed mainly through safety engineers in mining companies [7]. If the knowledge that they want to teach through training or induction is not transmitted adequately, the most valuable resource of companies, that is the human, would be more exposed to danger due to ignorance about the topics of safety, highlighting the great importance of education in the field of engineering.

The world's leading mining companies advise providing experiential learning in inductions and training through new interactive and immersive technologies such as podcasting, social media, interactive whiteboards, and virtual reality [8]. A study [9] examining the positive impact of the digital learning tool EPIC on 14,000 people found that the actor-based teaching application generates a more interactive and immersive pedagogical experience than a conventional classroom. It is also necessary to specify that a study [10] that used high fidelity simulation (HFS) identified that carrying out simulations more frequently and having the opportunity to "repeat" the experience would improve student learning. It is essential to mention that training, technical, and financial support requires the mining company's attention to implementing digital learning tools and achieving success in technology-supported education.

Technology acceptance model 3 (TAM3) was developed by Venkatesh [11]. It is essential to know the participants' opinions to analyze whether a new learning tool will be implemented and properly used, bringing considerable benefits [4] for students, teacher trainers, and the company where the pedagogical innovation would be applied. It has been researched with many different external variables in the literature [12]; it is an information systems theory that models how users come to accept and use technology.

The main objective of this work is to conduct a pilot test for practitioners and interns of mining companies in southern Peru to know their acceptance of digital learning tools using the TAM3.

II. RELATED WORK

Different works have been reviewed on the proposed theme that is shown below.

A. Digital Learning Tools

Xu, Zhang, and Hou, in their paper [13], indicate that the traditional teaching methods such as 2D presentations, conferences, and workshops have been innovated by technological learning tools such as simulations, immersive virtual reality, augmented and mixed reality, mobile devices, online training platforms, and game engine techniques, among others. The importance of learning tools lies in improving teaching quality to meet external challenges and demands in education [14].

1) *Benefits for students and teacher trainers:* The use of digital learning tools would allow the student to obtain more virtual practical experience, being able to make mistakes since it is considered a safe place, and thus improve their performance and dexterity when facing an actual situation [10], solving problems with confidence and without anxiety. This benefit would enhance the student's emotional attitude, making them feel motivated and satisfied with learning in a non-traditional way [15]. In the paper by Kurniawan, Suharjito, Diana, and Witjaksono [15], it is worth mentioning that using these tools allows the student to assimilate the information better when dealing with complex subjects to understand. It also creates opportunities for discussion within the classroom since digital learning tools allow students to play an active role, unlike conventional classes, where they play a passive role [16].

Digital teaching tools provide the teacher with better pedagogical technologies than traditional methods to explain and make their students understand, thanks to these tools' interactivity capacity, making the learning process more attractive, fun, and collaborative [17]. These results encourage teachers to continue using innovative learning tools with their students [3], increasing teacher effectiveness, increasing the quality of education, reducing the burden of teaching complicated subjects, and improving student performance.

B. Safety Culture and the Peruvian Regulation of Safety and Occupational Health in Mining

Mandhani, Nayak, and Parida [18] mention that safety is essential for all companies, especially mining companies, so the structure and culture of security are critical elements in safeguarding workers' lives.

The development of a safety culture in a mining company is essential to improve the rule-following behavior of employees to avoid the existence of incidents or accidents [19]. It is an indispensable medium for ensuring good performance in safety matters and significantly reducing accidents [20] [21]. According to Zhao, Zhao, Davidson, and Zuo [22], the economic loss due to occupational accidents represents approximately 8.5% of the project costs. For this reason, companies could be interested in providing safe working conditions to avoid expenses and safeguard human lives, where unique pedagogical campaigns such as training, inductions,

health promotion, risk assessment, and the environment can be offered [23]. In Peru, mining companies are governed under the D.S. N ° 023-2017-EM [24] in terms of occupational health and safety, article 72 specifies the minimum number of hours of necessary induction (eight (8) hours during four (4) days in mining activities of high risk and eight (8) hours for two (2) days in lower risk mining activities). There, an entry to the mining operation must carry out very apart from the necessary induction; the number of courses may vary according to the area to enter. In the required induction, topics such as occupational health and safety policies of the company, Internal Regulation of Safety and Occupational Health, Traffic Rules, High-risk jobs, emergency plans in the mining company, and definitions of danger and risk must be learned effectively. Risk controls, among others. So, trainers must be committed to the quality of learning that students will receive to ensure compliance with the law and safeguard lives.

C. Technological Acceptance Model (TAM3)

The TAM3 model was developed by Venkatesh [11] to determine the perceived usefulness and ease of use of the technology. The paper of Dönmez-Turan and Kir [25] indicates that the TAM3 was used to define the adoption of practitioners and interns working in mining companies in southern Peru to use new digital learning tools in safety inductions before entering the mine. As is shown in Table I, all the variables of the TAM3 used in the present investigation are defined. There is currently not enough research about implementing new digital learning tools in training and induction in mining companies. Since applying these tools in schools or universities is more studied, leaving aside the attention that industrial companies need in terms of education in security [26].

TABLE I. VARIABLES OF THE TAM3

Variable	Conceptualization
Image (IMG)	The level at which an individual feels that technological innovation will increase their social status [26].
Job Relevance (REL)	The degree to which a person believes that technological innovation can be applied to their work.
Output Quality (OUT)	The extent to which an individual believes that technological innovation can have a correct functionality in their work activities.
Result Demonstrability (RES)	The level at which a person thinks that technological innovation results can be observed and communicated.
Perceived Usefulness (PU)	The extent to which an individual perceives that technological innovation will improve their performance.
Computer Anxiety (CANX)	The degree to which an individual feels fear and insecurity when using technological innovations.
Computer Playfulness (CPLAY)	Level of enthusiasm that a person has when using technological innovation.
Computer Self-efficacy (CSE)	Measure which an individual feels capable of manipulating a technological innovation to carry out their activities.
Perceived Enjoyment (ENJ)	Level of enjoyment and satisfaction when using technological innovation.

Variable	Conceptualization
Objective Usability (OU)	The measure indicates how easy or difficult it will be to carry out a specific activity with technological innovation.
Perceptions of External Control (PEC)	The degree to which a person thinks there is adequate support to provide technical assistance for technological innovation.
Subjective Norm (SN)	Level of influence of one individual over another for the use of technological innovation.
Experience (EXP)	The extent to which an individual uses a technological innovation [11].
Voluntariness (VOL)	The degree to which a person would use a technological innovation without being forced [26].
Perceived Ease of Use (PEOU)	Measure how a person feels and how easy it will be to use technological innovation.
Behavioral Intention (BI)	Level of intention that an individual has to use technological innovation.
Use Behavior (USE)	Degree of permanence of a person in the face of technological innovation.

III. METHODOLOGY

A. Participants

This study was directed to 46 practitioners and interns from different mining companies in Southern Peru. They are between 20 and 30 years old; the majority are between 22 and 25 years old, representing 85% of the sample. The survey was shared using Google Forms [27]; so that it could have reached each participant by overcoming any territorial limitations. Next, Table II shows the distribution of survey participants; they were divided between people who work in surface mining (open pit) or underground mining (sinkhole).

TABLE II. DISTRIBUTION OF SURVEY PARTICIPANTS

Type of field	Number of participants	Percentage
Surface or open-pit mining	42	91%
Underground or sinkhole mining	4	9%
Total	46	100%

B. TAM3 Measuring Instrument

The Technological Acceptance Model (TAM3) proposed by Venkatesh was used. A network of variables was developed for adoption, information technologies use and empirically tested in the proposed integrated model, focusing on the application of TAM3 both a priori and after implementing the technology. It was shown that TAM3 provides necessary help in making managerial decisions to implement new information technologies [11]. The TAM primarily analyzes the technological acceptance of innovative methods and tools in education. It was shown to be an excellent instrument for understanding the participants' intentions [16]. We decided to use the TAM3 over other essential tools in the measurement of technological acceptance, such as the Unified Theory of

Acceptance and Use of Technology 2 (UTAUT2), since the latter includes variables such as age, gender, and price [28], whose parameters were considered not necessary for this research. A survey was conducted, taking into account the considerations of each TAM3 variable. They are defined in Table I. It was decided to use the Likert scale to establish a multiple-choice response range [29], having a range from 1 up to 7 points, where one meant "Strongly Disagree" and seven told, "Strongly Agree." Table III shows the questionnaire questions carried out on January 24, 2022.

TABLE III. QUESTIONNAIRE VARIABLES

Variable	Questionnaire
Image (IMG)	Receiving induction classes using learning technologies raises the status of the mining company where I work.
Job Relevance (REL)	I feel that using learning technologies in induction classes will positively impact the various tasks related to my work.
Output Quality (OUT)	Learning technologies could have excellent results in my understanding of different induction topics.
Result Demonstrability (RES)	It could explain why learning technologies can be beneficial in understanding the information received in induction classes.
Perceived Usefulness (PU)	I think it is possible that the use of new learning technologies in my induction classes would improve my job performance.
Computer Anxiety (CANX)	I feel safe and comfortable using new learning technologies.
Computer Playfulness (CPLAY)	I am excited about using new learning technologies in my induction classes.
Computer Self-efficacy (CSE)	I have the necessary skills to use new learning technologies without difficulty in induction classes.
Perceived Enjoyment (ENJ)	It would be more fun to use new learning technologies in my induction classes at the mining company.
Objective Usability (OU)	I think the new learning technologies would be easy and dynamic to use.
Perceptions of External Control (PEC)	Having the necessary opportunities, knowledge, and resources, it would be easy for me to use new learning technologies.
Subjective Norm (SN)	I think induction teachers think that new learning technologies could be helpful.
Experience (EXP)	I have experience in the use of a technological learning system.
Voluntariness (VOL)	I think that I would voluntarily access my induction classes with new learning technologies.
Perceived Ease of Use (PEOU)	I believe that new learning technologies will be easy to use.
Behavioral Intention (BI)	If I have the opportunity to take induction classes using new learning technologies, I would be happy to try it.
Use Behavior (USE)	I would use new learning technologies frequently in my induction classes.

C. Statistical Analysis with PLS-SEM

The obtained results in Google Forms were exported to an Excel file, delimited by commas. Finally, the Smart-PLS 3 software was used to obtain, through the partial least squares technique in structural equation models (PLS-SEM), statistical data such as minimum and maximum responses for each variable, standard deviations, the correlation coefficient between variables, coefficient route, hypothesis t-test, frequencies, percentages, arithmetic means, and discriminant validity [30].

IV. RESULTS

In Table IV, we observed that all the arithmetic means of the responses of each variable of the Technological Acceptance Model exceeded the neutral state given in the Likert scale (4 - "Neither disagree nor agree"). There is also highlighted that all the questions had the highest level on the Likert scale had a maximum score (7 - "Totally agree"), so we can affirm that the surveyed participants have the intention and acceptance of the use of new technology learning tools to be implemented in safety induction classes before entering the mining company. The variable with the highest average response was (BI), with a value of 6,326 on the Likert scale; this was followed by (USE) with a value of 6,196 on the same scale, reaffirming the technological acceptance by participants for the implementation of mining educational tools. Likewise, these last two were the variables with the lowest standard deviation, having a value of 0.957 and 0.947, respectively, which means the homogeneity of the surveyed responses. The variables with the lowest average response were Experience (EXP) and Subjective Norm (SN), with values of 5.500 and 5.674, respectively; in the same way, these two variables had the highest standard deviation with values of 1.514 and 1.445, respectively, demonstrating the high dispersion of responses from each intern and practitioner, having a minimum of 1 and a maximum of 7 on the Likert scale. It should be mentioned [11] that these are two of the most critical variables of the TAM3.

The bilateral correlation of the TAM3 variables is presented in Table V. It shows all the correlations have a strong and positive correlation, indicating coherence between variables. In Table VI, we appreciated to determine if the constructs have discriminant validity, a statistical test used to demonstrate the multicollinearity of the proposed model; this should be addressed if certain variables exceed the threshold for possible multicollinearity (0.7) [31]. We explained that the model has good reliability and construct validity.

In Fig. 1, the standardized trajectory coefficients between all the variables performed by Venkatesh [26] and the substantial variance (R²) in each of its four endogenous variables [32] can be appreciated. It is essential to mention that only four hypotheses were validated for the study according to the t-test where the value of 1.96 has to be exceeded [33], these being the relationships between IMG -> PU, REL -> PU, SN -> IMG, and BI -> USE, 2.422, 2.174, 4.139 and 5.488 respectively.

TABLE IV. DESCRIPTIVE STATISTICS

	Nº	Mean	Min	Max	Standard deviation
PU	1	5.870	2	7	1.172
IMG	2	6.000	2	7	1.123
REL	3	5.783	2	7	1.301
OUT	4	6.000	3	7	1.043
RES	5	5.761	4	7	1.087
PEOU	6	5.804	2	7	1.262
CANX	7	6.022	4	7	1.011
CPLAY	8	6.065	3	7	1.009
CSE	9	6.152	4	7	1.042
ENJ	10	5.978	1	7	1.242
OU	11	5.848	3	7	1.179
PEC	12	6.152	4	7	0.999
SN	13	5.674	1	7	1.445
EXP	14	5.500	1	7	1.514
VOL	15	6.109	2	7	1.184
BI	16	6.326	4	7	0.957
USE	17	6.196	4	7	0.947

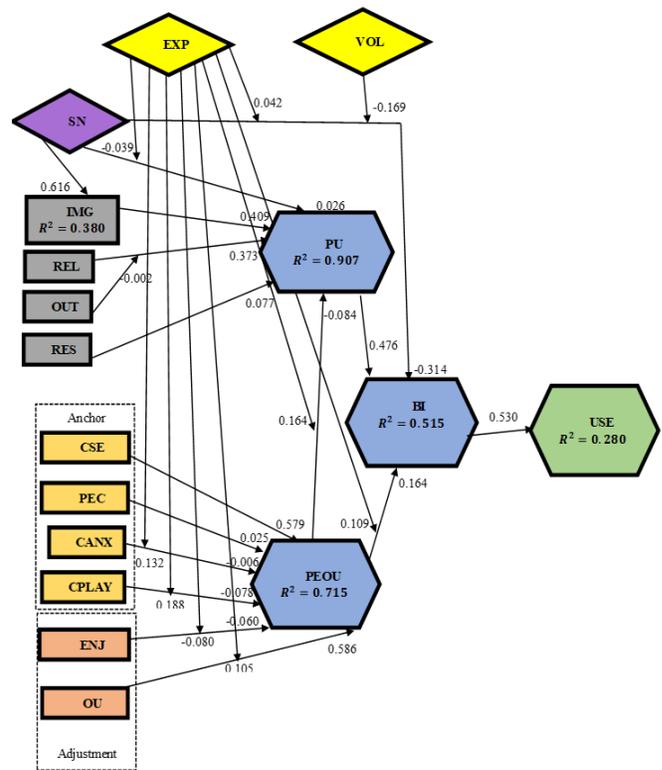


Fig. 1. Path Coefficients and Variances are Explained (R²). Where, Image (IMG), Job Relevance (REL), Output Quality (OUT), Result Demonstrability (RES), Perceived Usefulness (PU), Computer Anxiety (CANX), Computer Playfulness (CPLAY), Computer Self-efficacy (CSE), Perceived Enjoyment (ENJ), objective Usability (OU), Perceptions of External Control (PEC), Subjective Norm (SN), Experience (EXP), Voluntariness (VOL), Perceived Ease of Use (PEOU), Behavioral Intention (BI).

TABLE V. BILATERAL CORRELATIONS

	<i>PU</i>	<i>IMG</i>	<i>REL</i>	<i>OUT</i>	<i>RES</i>	<i>PEOU</i>	<i>CANX</i>	<i>CPLAY</i>	<i>CSE</i>	<i>ENJ</i>	<i>OU</i>	<i>PEC</i>	<i>SN</i>	<i>EXP</i>	<i>VOL</i>	<i>BI</i>	<i>USE</i>
<i>PU</i>	1.000																
<i>IMG</i>	0.892	1.000															
<i>REL</i>	0.851	0.729	1.000														
<i>OUT</i>	0.729	0.706	0.561	1.000													
<i>RES</i>	0.709	0.588	0.778	0.595	1.000												
<i>PEOU</i>	0.438	0.414	0.557	0.463	0.521	1.000											
<i>CANX</i>	0.498	0.460	0.417	0.475	0.638	0.515	1.000										
<i>CPLAY</i>	0.761	0.768	0.558	0.765	0.490	0.386	0.489	1.000									
<i>CSE</i>	0.319	0.297	0.281	0.380	0.454	0.618	0.740	0.322	1.000								
<i>ENJ</i>	0.730	0.748	0.603	0.537	0.592	0.427	0.381	0.712	0.322	1.000							
<i>OU</i>	0.678	0.542	0.673	0.548	0.650	0.711	0.477	0.575	0.514	0.710	1.000						
<i>PEC</i>	0.648	0.542	0.644	0.668	0.534	0.541	0.147	0.529	0.291	0.581	0.647	1.000					
<i>SN</i>	0.655	0.616	0.633	0.476	0.504	0.418	0.258	0.671	0.177	0.710	0.468	0.561	1.000				
<i>EXP</i>	0.245	0.268	0.221	0.124	0.376	0.347	0.533	0.278	0.710	0.399	0.481	0.165	0.253	1.000			
<i>VOL</i>	0.527	0.442	0.510	0.617	0.628	0.334	0.452	0.540	0.427	0.445	0.355	0.611	0.402	0.067	1.000		
<i>BI</i>	0.600	0.445	0.703	0.436	0.639	0.431	0.600	0.451	0.386	0.335	0.526	0.448	0.328	0.218	0.564	1.000	
<i>USE</i>	0.767	0.675	0.688	0.573	0.573	0.360	0.314	0.761	0.256	0.688	0.669	0.658	0.682	0.235	0.582	0.530	1.000

TABLE VI. DISCRIMINANT VALIDITY

	<i>BI</i>	<i>CANX</i>	<i>CPLAY</i>	<i>CSE</i>	<i>ENJ</i>	<i>EXP</i>	<i>IMG</i>	<i>OU</i>	<i>OUT</i>	<i>PEC</i>	<i>PEOU</i>	<i>PU</i>	<i>REL</i>	<i>RES</i>	<i>SN</i>	<i>USE</i>	<i>VOL</i>
<i>BI</i>																	
<i>CANX</i>	0.600																
<i>CPLAY</i>	0.451	0.489															
<i>CSE</i>	0.386	0.740	0.322														
<i>ENJ</i>	0.335	0.381	0.712	0.322													
<i>EXP</i>	0.218	0.533	0.278	0.710	0.399												
<i>IMG</i>	0.445	0.460	0.768	0.297	0.748	0.268											
<i>OU</i>	0.526	0.477	0.575	0.514	0.710	0.481	0.542										
<i>OUT</i>	0.436	0.475	0.765	0.380	0.537	0.124	0.706	0.548									
<i>PEC</i>	0.448	0.147	0.529	0.291	0.581	0.165	0.542	0.647	0.668								
<i>PEOU</i>	0.431	0.515	0.386	0.618	0.427	0.347	0.414	0.711	0.463	0.541							
<i>PU</i>	0.600	0.498	0.761	0.319	0.730	0.245	0.892	0.678	0.729	0.648	0.438						
<i>REL</i>	0.703	0.417	0.558	0.281	0.603	0.221	0.729	0.673	0.561	0.644	0.557	0.851					
<i>RES</i>	0.639	0.638	0.490	0.454	0.592	0.376	0.588	0.650	0.595	0.534	0.521	0.709	0.778				
<i>SN</i>	0.328	0.258	0.671	0.177	0.710	0.253	0.616	0.468	0.476	0.561	0.418	0.655	0.633	0.504			
<i>USE</i>	0.530	0.314	0.761	0.256	0.688	0.235	0.675	0.669	0.573	0.658	0.360	0.767	0.688	0.573	0.682		
<i>VOL</i>	0.564	0.452	0.540	0.427	0.445	0.067	0.442	0.355	0.617	0.611	0.334	0.527	0.510	0.628	0.402	0.582	

V. DISCUSSION

Our results show that interns' and practitioners' technological acceptance were successful since they are willing to adapt to new digital learning tools, such as in research where a group of students intends to use new pedagogical strategies [4]. They could find the implementation and functionalities interesting, supporting their knowledge, as happened in a study

that adopted an interactive web environment technology called MAgAdI at the University of the Basque Country [3]. It is intended to give the initiative to implement new digital learning tools in the mining industry, as recommended by a study where its findings foster crucial management decisions for the future implementation of information technologies [11]. It helps improve the academic performance effectively in the induction classes of the interns and practitioners before

entering the mine, as is the case of a study that provided facilities of its results to school providers to use new innovative pedagogical methods [34]. It is necessary to highlight that the topics offered in the induction classes of the participants are both theoretical and practical so that the new learning tools could contribute to the improvement of their performance, as argued in a study where the Perception of Mining Engineering students, concluding that the use of technology in education is necessary to improve student performance [4]. Learning tools allow the user to be transferred to a virtual environment where it is allowed to make mistakes and learn through experience safely, as indicated in an investigation with simulated immersive learning platforms where their findings indicated that students could avoid taking risks to demonstrate their capabilities in a safe environment [10]. A study conducted in 2019 mentioned that virtual laboratories might have the same or better learning outcomes than laboratories traditional [16], providing a unique learning experience [3], breaking the conventional scheme, and overlapping conventional learning [6], as we stated in the present study. In the present study, 46 pilot samples of different interns and practitioners of mining companies located in the macro-southern region of Peru were investigated to determine technological acceptance through TAM3, as was done in an article [35] where 44 respondents were studied, specifying that no there would be a problem in working with a small sample size since the results will be stable.

As we mentioned in the survey instructions, by new learning technologies, we mean: "virtual or augmented reality, holograms, interactive platforms, and mobile applications" according to the results had a positive impact on all TAM3 variables, as formulated by Venkatesh [35], especially in Use Behavior (USE) and Behavioral Intention (BI). The same situation occurred in a scientific production that shows that students' perception of virtual laboratories, simulators, interactive learning activities, and game-based learning positively influenced satisfaction, usefulness, and perceived ease of use [16]. An additional question was asked to the TAM3 to determine the preference of the studied participants towards a conventional class or a class with digital learning tools, where 89.13% of the total participants chose classes with innovative learning tools since these can keep employees focused positively and proactively as mentioned in a study conducted in Australia [23]. The participants in the presented studies rejected a conventional class, defined as a "lesson given by a teacher or trainer using 2D tools, such as slides, whiteboard, flipcharts, non-interactive videos"; it can generally cause stress and anxiety, reducing the value of learning [10] and being an obstacle to the adoption of a digital learning tool [25]. The present study used the partial least squares technique in structural equation models (PLS-SEM). Being very powerful and not having a minimum sample parameter gives us extra support for the validity of the results, as Said Al-Gahtani comments [32]. SmartPLS 3 software analyzed these data; it was used from the same firm by different studies associated with TAM evaluation [36] [37]. The results of the discriminant variance correlations mostly exceed the threshold of 0.7, confirming the construct's validity and the trajectory correlations of the present study. These were positive and strong, as evidenced by studies conducted between 2015 and

2021 [31][38]. The results also indicate that IMG significantly influences PU, as happens in a study carried out in Saudi Arabia, where it is specified that social influence processes were shown to affect perceived profit. On the other hand, the same study found that the instrumental cognitive function positively impacts Perceived Utility, as this article [32] did when validating the REL -> PU hypotheses.

It has been considered to use all external variables of Technology Acceptance Model 3 because they are considered more relevant concerning Unified Theory of Acceptance and Use of Technology 2 (UTAUT2), one of the essential acceptance models TAM3 [1]. The TAM3 was considered, unlike UTAUT2, because the external variables of the TAM3 facilitate the intention and maintain voluntariness [24], unlike UTAUT2, which predominates individual conditions such as gender, age, price value.

VI. CONCLUSION

Safety is necessary to provide qualified education to collaborators to increase their effectiveness by voluntary rules compliance, preserving the human resource that is the most crucial element of the companies, and saving accident expenses. It can be achieved using more effective pedagogical tools than traditional ones. However, it is essential to know if staff are willing to accept new learning technologies such as virtual or augmented reality, holograms, interactive platforms, and mobile applications. Before joining a mining company, interns and practitioners intend to use new technological learning tools in induction classes. We recommended expanding the sample of hired workers for future research work since they receive constant safety talks specified by the D.S. N ° 023-2017-EM, since integral management of the safety culture is being promoted, that goes from the highest rank to the lowest level; consequently, this would cause a change in attitude in the company members, making them take care of each other. We recommend mining and industrial companies implement new digital learning tools to train human resources and improve the development of their skills in occupational safety and health. There is limited research on technological acceptance to help management decision-making. It is necessary to imply that studies on the implementation of pedagogical technologies should not be centralized only in schools or universities, mining, and industrial companies where education can save lives, so we recommended developing scientific production in that theme and domain.

REFERENCES

- [1] J. W. M. Lai and M. Bower, "How is the use of technology in education evaluated? A systematic review," *Comput. Educ.*, vol. 133, 2019, doi: 10.1016/j.compedu.2019.01.010.
- [2] S. O'Connor and T. Andrews, "Smartphones and mobile applications (apps) in clinical nursing education: A student perspective," *Nurse Educ. Today*, vol. 69, 2018, doi: 10.1016/j.nedt.2018.07.013.
- [3] A. Álvarez, M. Martín, I. Fernández-Castro, and M. Urretavizcaya, "Blending traditional teaching methods with learning environments: Experience, cyclical evaluation process and impact with MAgAdI," *Comput. Educ.*, vol. 68, 2013, doi: 10.1016/j.compedu.2013.05.006.
- [4] P. López, J. Rodríguez, A. Acosta, and M. Berrios, "Analysis from the student perspective on the implementation of learning technologies in mining engineering," in *CEUR Workshop Proceedings*, 2019, vol. 2555.

- [5] M. Kangas, P. Siklander, J. Randolph, and H. Ruokamo, "Teachers' engagement and students' satisfaction with a playful learning environment," *Teach. Teach. Educ.*, vol. 63, 2017, doi: 10.1016/j.tate.2016.12.018.
- [6] C. wen Shen and J. tsung Ho, "Technology-enhanced learning in higher education: A bibliometric analysis with latent semantic approach," *Comput. Human Behav.*, vol. 104, 2020, doi: 10.1016/j.chb.2019.106177.
- [7] A. Adyatama, D. H. Syaifullah, and B. N. Moch, "Evaluating the role of cognitive factors on hauling workplace accidents: Study case at Central Borneo coal mining company," in *AIP Conference Proceedings*, 2020, vol. 2227, doi: 10.1063/5.0007247.
- [8] J. Zhang, J. Fu, H. Hao, G. Fu, F. Nie, and W. Zhang, "Root causes of coal mine accidents: Characteristics of safety culture deficiencies based on accident statistics," *Process Saf. Environ. Prot.*, vol. 136, 2020, doi: 10.1016/j.psep.2020.01.024.
- [9] E. J. Harvey, J. A. Pinder, R. A. Haslam, A. R. J. Dainty, and A. G. Gibb, "The use of actor-based immersive health and safety inductions: Lessons from the Thames Tideway Tunnel megaproject," *Appl. Ergon.*, vol. 82, 2020, doi: 10.1016/j.apergo.2019.102955.
- [10] H. MacLean, K. J. Janzen, and S. Angus, "Lived Experience in Simulation: Student Perspectives of Learning From Two Lenses," *Clin. Simul. Nurs.*, vol. 31, 2019, doi: 10.1016/j.ecns.2019.03.004.
- [11] V. Venkatesh and H. Bala, "Technology acceptance model 3 and a research agenda on interventions," *Decis. Sci.*, vol. 39, no. 2, 2008, doi: 10.1111/j.1540-5915.2008.00192.x.
- [12] A. Donmez-Turan and M. T. Odabas, "Evaluating Technology Acceptance Model on the User Resistance Perspective: A Meta-analytic Approach," in *Communications in Computer and Information Science*, 2022, vol. 1534 CCIS, doi: 10.1007/978-3-030-96040-7_59.
- [13] S. Xu, M. Zhang, and L. Hou, "Formulating a learner model for evaluating construction workers' learning ability during safety training," *Saf. Sci.*, vol. 116, 2019, doi: 10.1016/j.ssci.2019.03.002.
- [14] Q. Liu, S. Geertshuis, and R. Grainger, "Understanding academics' adoption of learning technologies: A systematic review," *Comput. Educ.*, vol. 151, 2020, doi: 10.1016/j.compedu.2020.103857.
- [15] J. M. Zydney and Z. Warner, "Mobile apps for science learning: Review of research," *Comput. Educ.*, vol. 94, 2016, doi: 10.1016/j.compedu.2015.11.001.
- [16] R. Estriegana, J. A. Medina-Merodio, and R. Barchino, "Student acceptance of virtual laboratory and practical work: An extension of the technology acceptance model," *Comput. Educ.*, vol. 135, 2019, doi: 10.1016/j.compedu.2019.02.010.
- [17] M. H. Kurniawan, Suharjito, Diana, and G. Witjaksono, "Human Anatomy Learning Systems Using Augmented Reality on Mobile Application," in *Procedia Computer Science*, 2018, vol. 135, doi: 10.1016/j.procs.2018.08.152.
- [18] J. Mandhani, J. K. Nayak, and M. Parida, "Interrelationships among service quality factors of Metro Rail Transit System: An integrated Bayesian networks and PLS-SEM approach," *Transp. Res. Part A Policy Pract.*, vol. 140, 2020, doi: 10.1016/j.tra.2020.08.014.
- [19] E. Stemn, C. Bofinger, D. Cliff, and M. E. Hassall, "Examining the relationship between safety culture maturity and safety performance of the mining industry," *Saf. Sci.*, vol. 113, 2019, doi: 10.1016/j.ssci.2018.12.008.
- [20] H. Lu and H. Chen, "Does a people-oriented safety culture strengthen miners' rule-following behavior? The role of mine supplies-miners' needs congruence," *Saf. Sci.*, vol. 76, 2015, doi: 10.1016/j.ssci.2015.02.018.
- [21] M. Rubin, A. Giacomini, R. Allen, R. Turner, and B. Kelly, "Identifying safety culture and safety climate variables that predict reported risk-taking among Australian coal miners: An exploratory longitudinal study," *Saf. Sci.*, vol. 123, 2020, doi: 10.1016/j.ssci.2019.104564.
- [22] Z. Y. Zhao, X. J. Zhao, K. Davidson, and J. Zuo, "A corporate social responsibility indicator system for construction enterprises," *J. Clean. Prod.*, vol. 29–30, 2012, doi: 10.1016/j.jclepro.2011.12.036.
- [23] A. M. Vecchio-Sadus and S. Griffiths, "Marketing strategies for enhancing safety culture," *Saf. Sci.*, vol. 42, no. 7, 2004, doi: 10.1016/j.ssci.2003.11.001.
- [24] D.S. N° 023-2017-EM, "Reglamento de Seguridad y Salud Ocupacional en Minería D.No 023-2017-EM," *D. Of. El Peru.*, 2017.
- [25] A. Dönmez-Turan and M. Kir, "User anxiety as an external variable of technology acceptance model: A meta-analytic study," in *Procedia Computer Science*, 2019, vol. 158, doi: 10.1016/j.procs.2019.09.107.
- [26] M. Loosemore and N. Malouf, "Safety training and positive safety attitude formation in the Australian construction industry," *Saf. Sci.*, vol. 113, 2019, doi: 10.1016/j.ssci.2018.11.029.
- [27] T. L. Wiemken, S. P. Furmanek, W. A. Mattingly, J. Haas, J. A. Ramirez, and R. M. Carrico, "Googling your hand hygiene data: Using Google Forms, Google Sheets, and R to collect and automate analysis of hand hygiene compliance monitoring," *Am. J. Infect. Control*, vol. 46, no. 6, 2018, doi: 10.1016/j.ajic.2018.01.010.
- [28] V. Venkatesh, J. Y. L. Thong, and X. Xu, "Consumer acceptance and use of information technology: Extending the unified theory of acceptance and use of technology," *MIS Q. Manag. Inf. Syst.*, vol. 36, no. 1, 2012, doi: 10.2307/41410412.
- [29] S. E. Harpe, "How to analyze Likert and other rating scale data," *Currents in Pharmacy Teaching and Learning*, vol. 7, no. 6, 2015, doi: 10.1016/j.cptl.2015.08.001.
- [30] M. Martínez Ávila and E. Fierro Moreno, "Aplicación de la técnica PLS-SEM en la gestión del conocimiento: un enfoque técnico práctico / Application of the PLS-SEM technique in Knowledge Management: a practical technical approach," *RIDE Rev. Iberoam. para la Investig. y el Desarro. Educ.*, vol. 8, no. 16, 2018, doi: 10.23913/ride.v8i16.336.
- [31] K. M. S. Faqih and M. I. R. M. Jaradat, "Assessing the moderating effect of gender differences and individualism-collectivism at individual-level on the adoption of mobile commerce technology: TAM3 perspective," *J. Retail. Consum. Serv.*, vol. 22, 2015, doi: 10.1016/j.jretconser.2014.09.006.
- [32] S. S. Al-Gahtani, "Empirical investigation of e-learning acceptance and assimilation: A structural equation model," *Appl. Comput. Informatics*, vol. 12, no. 1, 2016, doi: 10.1016/j.aci.2014.09.001.
- [33] S. A. Kamal, M. Shafiq, and P. Kakria, "Investigating acceptance of telemedicine services through an extended technology acceptance model (TAM)," *Technol. Soc.*, vol. 60, 2020, doi: 10.1016/j.techsoc.2019.101212.
- [34] D. R. Compeau and C. A. Higgins, "Computer self-efficacy: Development of a measure and initial test," *MIS Q. Manag. Inf. Syst.*, vol. 19, no. 2, 1995, doi: 10.2307/249688.
- [35] V. Venkatesh and F. D. Davis, "Theoretical extension of the Technology Acceptance Model: Four longitudinal field studies," *Manage. Sci.*, vol. 46, no. 2, 2000, doi: 10.1287/mnsc.46.2.186.11926.
- [36] P. R. Schulman, "Organizational structure and safety culture: Conceptual and practical challenges," *Saf. Sci.*, vol. 126, 2020, doi: 10.1016/j.ssci.2020.104669.
- [37] M. Ramkumar, T. Schoenherr, S. M. Wagner, and M. Jenamani, "Q-TAM: A quality technology acceptance model for predicting organizational buyers' continuance intentions for e-procurement services," *Int. J. Prod. Econ.*, vol. 216, 2019, doi: 10.1016/j.ijpe.2019.06.003.
- [38] J. C. Sánchez-Prieto, S. Olmos-Migueláñez, and F. J. García-Peñalvo, "MLearning and pre-service teachers: An assessment of the behavioral intention using an expanded TAM model," *Comput. Human Behav.*, vol. 72, 2017, doi: 10.1016/j.chb.2016.09.061.

Enhanced Symbol Recognition based on Advanced Data Augmentation for Engineering Diagrams

Ong Kai Bin¹, Yew Kwang Hooi², Said Jadid Abdul Kadir³, Haruhiro Fujita⁴, Luqman Hakim Rosli⁵

Department of Computer and Information Sciences, Universiti Teknologi PETRONAS, Malaysia^{1,2,3,5}
Faculty of Management and Information Sciences, Niigata University of International and Information Studies, Niigata City, Japan⁴

Abstract—Symbol recognition has generated research interest for image analytics of engineering diagrams. Techniques including structural, syntactic, statistical, Convolution Neural Network (CNN) were studied to identify gaps of research. Despite popularity, CNN requires huge learning dataset, which often involves costly procurement. To address this, combination between CycleGAN and CNN is proposed. CycleGAN generates more learning dataset synthetically, thus yielding opportunity to improve accuracy of symbol recognition. In the domain of for engineering symbols, standard CNN model is developed and used in experimental testing. Different ratios of training dataset were tested in multiple experiments using Piping and Instrument Diagram (P&IDs) drawings. Result of highest accuracy for symbol recognition is up to 92.85% against baseline and other method. The results determined that gradual reduction of training samples, the effectiveness of recognition accuracy performance after using proposed method was remained substantially stable.

Keywords—Symbol recognition; symbol spotting; engineering drawing; convolution neural network (CNN); CycleGAN; piping and instrument diagram (P&ID)

I. INTRODUCTION

Symbol recognition is a subset of ordinary pattern recognition which focusing on identifying, detecting, and recognizing components in technical drawings. Pattern recognition is a complicated process that requires to analyze data input, feature extraction, classification, and post processing. Therefore, various functions are needed for pattern recognizer. Recognizing familiar patterns automatically is an essential pattern. However, recognition does not work accurately when identifying and classifying unfamiliar objects. Insufficient data input is one of the factors.

An application of symbol recognition is for analysis of engineering drawing. Engineering drawings are frequently used in many fields such as Oil and Gas, manufacturing, construction, and engineering. In this study, Piping and Instrument Diagrams (P&ID) are selected. P&ID are schematic diagrams representing different components and flows in a manufacturing process design for a physical plant. These diagrams aid analysis of operability and safety of a process design.

More samples for engineering drawings may produce impactful benefits due to the application for plant safety. The technique is focusing on Oil and Gas field simultaneously can be generalized for other fields of engineering drawing too.

The motivation to study this technique is due to various factors:

- Current research & development trend focuses on smarter digital diagram for better image analytics. Application of the technology includes determining root cause and inferring risk of a safety deviation.
- Digitisation of schematic diagrams and image analytics are instrumental for Digital Twin (DT) in Cyber-Physical System (CPS) designs. DT provides richer information and more potentials than digital engineering drawing application. These include prospective planning, analysis of existing systems or process-parallel monitoring. In all cases, DT offers exceptional ability to create simulations in which various development and testing work can be carried out.

Identifying components inside a digital drawing is necessary in analysis. However, this is difficult because of the drawing's layout complexity. Wrong identification of any component can lead to a faulty analysis, thus posing risk to safety and operability.

There are three main types of pattern recognition mechanisms used to classify input data. Those types are statistical, structural (syntactic), neural. Statistical methods simply collect historical data and identify new patterns based on observations and analysis of that data. The structural technique is also known as the syntactic method since it is based on primitive sub-patterns. Machines do direct computing in the pattern recognition approaches covered so far. Mathematical and statistical techniques are used in direct calculations. Last but not least, neural approach applies biological concepts into technology for recognizing patterns. The result of this effort was the invention of artificial neural networks. A neural network is an information processing system. However, deep learning is becoming more popular as a result of its superior accuracy while training with huge volumes of data. A neural network-based deep learning method is used. Neural approach is less efficient at processing tasks than deep learning systems, which provide excellent efficiency and performance for tasks. Recently, Convolution Neural Network (CNN) is a class of deep neural networks with the most applications.

CNN has shortcomings for analysis of engineering drawings, in that the model used has not been sufficiently trained for engineering drawing domain. The challenge

concerns on huge quantity amount of data samples are required for training to obtain more accurate classification and recognition in training process. Some of the samples are difficult to collect for engineering drawings especially in P&ID diagrams [1]. The accuracy of recognition is important for analysis of engineering design issues such as operability and safety. Nonetheless, having to work hard to manually collect and correct a huge number of sample images for training remains a significant limitation[2]. Methods that rely on artificial training data are recommended. Therefore, a data augmentation technique is proposed called CycleGAN along with CNN to enhance the accuracy of symbol recognition.

The rest of this paper is organized as follows. Section II reviews the associated literatures with the research topic on pattern recognition approaches. Section III illustrates the proposed framework on enhancing the accuracy of recognition rate. Section IV presents the experimental results. Section V interprets the significant findings along with discussion. Section VI concludes the paper and mentions the future works in this study.

II. RELATED WORK

A. Symbol Recognition Trend

Symbol recognition and spotting is an innovative computer vision technology. This technology aims to replicate portions of the human visual system’s complexing, able to allow computers to recognize and interpret within images or videos. Engineering benefits mankind and technology symbolizes the future. Over the recent years, numerous challenges of computer vision were gradually emerged. Surprisingly, many

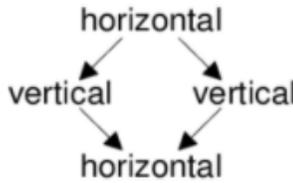
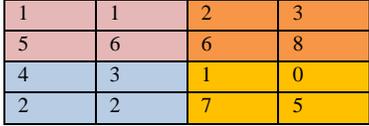
great efforts and solution have been cracked during the evolution timeline.

Table I summarizes the research trend for symbol spotting in the recent years. Symbol Spotting is an active topic in graphical symbol recognition and document analysis. In fact, some of the research papers are focusing on the sub-sections of the issues which are including symbol recognition, symbol detection, and primitive extraction. Those papers are altering from traditional symbol recognition towards the new conceptual of symbol recognition gradually. Symbol spotting has been enhanced in recent years persistently.

There is a demand to design an engineering drawings symbol recognition method compared to few approaches in Table I. Based on the review of history, the state-of-the-art approaches are keep replacing with better algorithms.

Statistical approach has been used precedent for pattern recognition since it is simple to manage. This approach mainly based on statistics and probabilities. Each pattern is obtained in terms of feature collection. The effectiveness of quality pattern depends on the set of feature extraction or measurement. It includes the pre-processing, and the segmentation is not required compared to local pattern representation. The decision boundary is established in the feature extraction via the analysis of probability distribution. Thus, different patterns allocate to the classes respectively. However, this approach has fault tolerant to image distortion since they tend to filter out small change in details.

TABLE I. EVOLUTION OF SYMBOL RECOGNITION TREND

Category approach	Pattern Recognition	Survey
Statistical: statistical features (geometric moments, R-signatures, etc.) + similarity measures	Number of segments: 4 Number of horizontal segments: 2 Number of vertical segments: 2 Number of diagonal segments: 0	Santosh et al. [3-5], 2015, 2016, 2018
Structural: graph-based matching + structural descriptors (visual primitives/ their spatial relations)		
Syntactic: inductive learning + programming spatial predicates (primitives’ relations)		
Semiautomatic and heuristic-based method in Convolutional Neural Network (CNN): Hough transform + connected component analysis + criterion	Pixel Values 	Elyan et al. [6], 2018
Clutter-tolerant cross-correlation for template matching		Rezvanifar et al. [7], 2019
Deep Learning Network in Convolutional Neural Network (CNN)		Yu et al. [8], 2019
Spatial Transformer Network (STN) in Convolutional Neural Network (CNN)		Y.Zhang[9], 2019
RPN method in R-CNN		Yun et al.[10], 2020
GNL method for both the small- and large-symbol networks		Kim et al.[11], 2021

In graphical documents, many objects can be graphically described, especially symbols [12]. Through structural recognition approach for symbols, the patterns can be represented as graphs. The characteristics of patterns are formed based on structure. Recognizing symbol is normally applied with structural recognition but this approach also can implement with other objectives such as learning, data structuring, indexing and others. It is also known as syntactic recognition occasionally since both are using the same formal language theory. Structural depends on the grammars to differentiate between data from several groups based on morphological interrelationships contained within the data. The sub patterns and the interactions between them that make up the data are represented by structural features, also known as primitives. However, recognize the primitives occur many troubles which primarily must focus on segmentation of noisy patterns and gram-mar inference from training data.

Syntactic recognition uses the structure of patterns and the syntax of language to construct a shape. The patterns are depended on the sentences in a language. The patterns are considered as sentences in a language, the primitives are viewed as the language's alphabet, and the sentences are formed using a grammar [12]. Thus, few numbers of primitives and grammatical rules are able to express complex patterns.

Structural and syntactic techniques are much closer to human comprehension. These techniques outperform the statistical in the case of complex symbols. However, the accuracy is obstructed the performance of the vectorization process, and matching based on graph representation which requires high computation cost.

Convolutional neural network (CNN) is a type of artificial neural network. This technique is the most popular to be applied in the recent year [8, 9, 13]. The employment of CNN is not restricted in image classification, recognition, and detection but also in Natural Language Processing (NLP). Therefore, the extensiveness of CNN application is broadly. The CNN achievements are also extravagance; produced many earth-shattering results in computer vision.

B. CNN Architecture in Image Processing

Most of the CNN architectures are based on supervised learning. A large amount of data is needed as a training sample to obtain more accurate classification and recognition in training process. Through enough training on data, CNN can overcome the limitations of the conventional methods such as inter-object interruption, morphological change, and noise problems [8]. However, some samples are difficult to collect. For example, the specific symbols in P&IDs drawings. It is extremely difficult to collect samples due to the limitations of the conditions.

CNN architecture model can be separated into two components which containing features extraction and classification. Feature extraction is performed by the convolution and pooling layers. Detect meaningful features in an image is a complicated task. The convolution layers have to learn such sophisticated features to illustrate the pattern shapes via pixels. Nan S. [14] experimented several feature extraction

algorithms at CNN model. FCN-CRF achieved highest average accuracy for image segmentation and average intersection-over-union compared with Support Vector Machine (SVM), K-means and FCN algorithms.

The few novelties [1, 6, 8] resulted the limited amount of training data for engineering drawing lower the accuracy of recognition. In additional, the lack of datasets occurs the class imbalance problem in classification.

C. Review Data Augmentation

Although more data samples can enhance the machine learning models, in fact collecting more data is not a best solution which channelling into a research topic. Data augmentation is a technique which providing more quantities of synthetic samples. This allows the algorithm to recognize and detect the specific components in an image precisely. Lan Goodfellow [15] proposed a framework called Generative Adversarial Network (GAN). The benefit of GAN can generate more synthetic samples to overcome the shortage and elevate the accuracy of object spotting. GAN is a type of machine learning model which consists of two neural networks compete with each other in order to provide more realistic image in the prediction.

GAN has been implemented in many applications previously for synthesizing more quality images and adding more training data in several studies. Shrivastava et al. [16] developed a GAN-based refiner network and enhanced the realism of simulated eye I mages by developing a GAN-based refiner network, resulting a 21% improvement in performance of an eye gaze estimation algorithm. In the field of medicine analyses, the shortage of image data is also commonly happened due to the lack of available images. Therefore, GAN play a role to synthesize realistic training data for liver lesion images [17], and brain MR images [18].

However, GAN training is highly unsteady because the discriminator and generator training needs to be delicately balanced. A common failure from mode collapse can be happened if the discriminator is too fierce or overwhelms the generator early during training, which results in convergence to a bad local optimum. Therefore, a new technique needs to be discovered instead of replacing GAN.

D. CycleGAN

CycleGAN is a subset of GAN technique which generating the automatic training of image-to-image translation models with unpaired examples. Some related works have been explored recently at the below.

Liu et. al. [19] applied stratified CycleGAN in medical images that generated graded variation in image quality. They resulted quality synthetic images using CycleGAN method. Zhang et. al. [20] presented a novel about road extraction method in generative adversarial network using Aerial images, require few samples and resulted better performance compared with several state-of-the-art techniques in term of detection accuracy. Park et. al. [21] implemented Dense-Net based framework in CycleGAN have eliminated data imbalance issue and deliver a better detection accuracy for wildfires images. Liu et.al. [22] equipped CycleGAN strategy to address ghosting problem. This work synthesizes video

context information and captures interframe stability better. Thus, CycleGAN is one of the impacts on the accuracy of recognition and spotting. The accuracy of the recognition can be influenced by producing more variance images.

III. METHODOLOGY

A. Symbols Dataset – Piping and Instrument Drawings (P&IDs)

Engineering drawings consists of several types of fields. Piping & Instrument Drawings is appointed to be our scope of study. The total number of collections from Piping & Instrument Drawings (P&IDs) is restricted into 7 sheets only due to limited public datasets available. However, this can highlight the problem then figure out a better solution.

In these experiments, only 7 types of symbols are extracted from the P&IDs. These types of symbols which including (a) Check valve, (b) Gate valve, (c) Gate_NC valve, (d) Globe valve, (e) Globe_NC valve, (f) Concentric reducer, (g) Weldcap. These symbols are shown in Fig. 1.

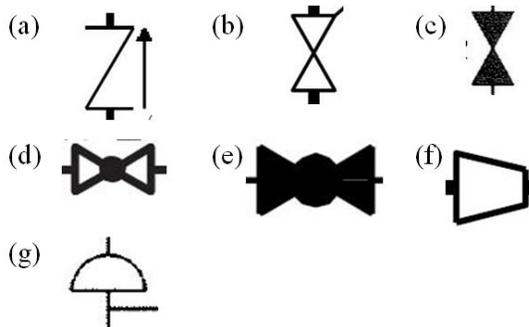


Fig. 1. Types of Symbols Extracted from P&IDs.

In additional, a total number collection of extracted symbols from P&IDs are 1,873. The following Fig. 2 summarized the number of symbols distributed for each class.

Each P&ID sheet displays different qualities of components. Qualities is one of the factors affects the accuracy for spotting. More data information with various qualities can provides the confidence to recognize the symbols precisely. Therefore, the limited symbols from P&IDs will conduct with advanced data augmentation technique to deliver more various qualities images.

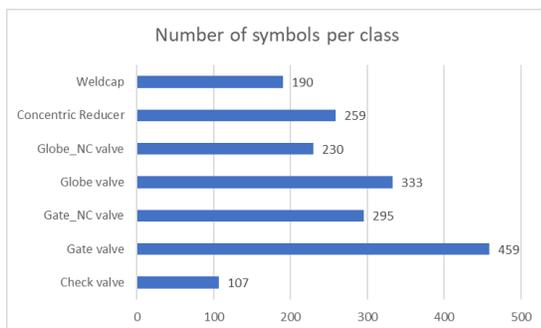


Fig. 2. Class Distribution of Symbols in Whole Dataset.

B. Symbol Recognition – CNN Algorithm

Symbols require an algorithm tool for recognizing. Therefore, a basic Convolution Neural Network (CNN) algorithm to be implemented. The implementation of CNN able to predict several class probabilities and bounding boxes simultaneously.

The CNN architecture involves several convolutional layers, max pooling layers and fully connected layers. A tensor which representing the data structure is required to proceed through several convolutional layers. Then, it is converted to a vector and then transmitted through a dense layer. The overview of CNN architecture is displayed in Fig. 3.

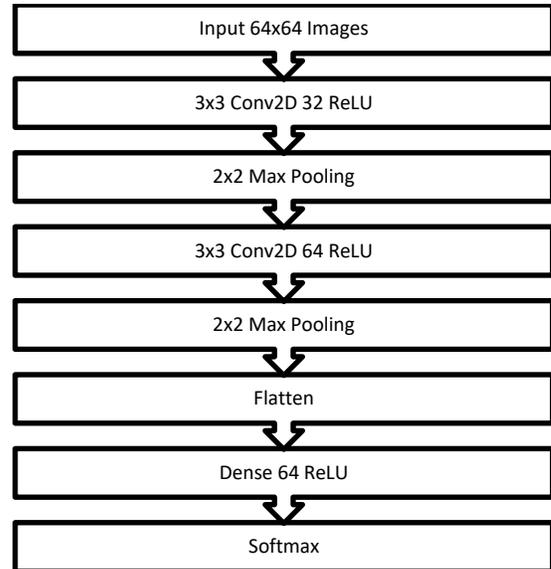


Fig. 3. Basic of CNN Architecture.

Input Images: All the extracted symbols from P&ID drawings and synthetics symbols are converted into 64x64 pixels.

$M \times M$ Conv2D N : In this basic CNN architecture, 2D convolution layer is used, this layer creates a convolution kernel or filter over the input data and perform an elementwise multiplication to produce a tensor of output pixel. Kernel is a convolution matrix or filter that used for features extraction in image processing including blurring, sharpening, embossing, edge detection, and others. $M \times M$ represents the size of filters, and N represents the numbers of filters that the convolution layer learns.

ReLU: Stands for Rectified Linear Unit. Commonly used as an activation function of the CNN layers and the fully connected layers because it is simple, fast, and empirically converge quickly and reliably when training a deep network.

$M \times M$ Max Pooling: Extract each patch from input featured maps. Generally, a filter of size 2x2 with a stride of 2 is implemented. The pooling is downsampling the feature map then spotlight the conspicuous feature in each patch.

Flatten: Also known as fully-connected layer for connecting the final classification model. Flatten the last output of feature map once the model has learnt the features. However, the flatten layer is used to convert the data into 1-dimensional array (1D) for the next layer's input.

Dense: A simple layer for neurons. Each neuron receives an input from the previous layer's neurons. This is used for classifying or predicting image based on the output from previous layers via shallow neural network.

Softmax: Commonly a final output layer in neural network. It is an activation function for generalization of the Sigmoid function but in multiple dimensions. It is used for performing multiclass classification and object recognition. Thus, it normalizes the output as probability distribution to each class.

C. Pre-processing

Prior to any work, the whole dataset containing different number of symbols for each class. All symbols were cropped into a size of 64 x 64 pixels.

According to previous related works, few novelties described the limitation of images. However, there is no findings specify an exact volume. Therefore, 3 methods were conducted in this study including Baseline CNN, CycleGAN + CNN, and Y.Zhang [9] algorithm as for comparison. Each method conducts several experiments with all the extracted symbols from P&IDs drawings. These symbols are classifying into three categories of datasets which including training dataset, validation dataset, and testing dataset. In these experiments, different ratios for training dataset are implemented while validation dataset and testing dataset remain unchanged. The highest ratio of training dataset is up to 4 and lowest ratio is 0.5. The scheme for classifying datasets and the description of each method is shown below:

In Fig. 4, the scheme illustrates the division for training dataset, validation dataset, and testing dataset. The ratio in whole dataset is 1:1:1. The training dataset is further using to implement on the synthetic models which described in Method 2 and Method 3 in the following paragraph. The synthetic images are generated via a model by applying the training dataset with different ratios.

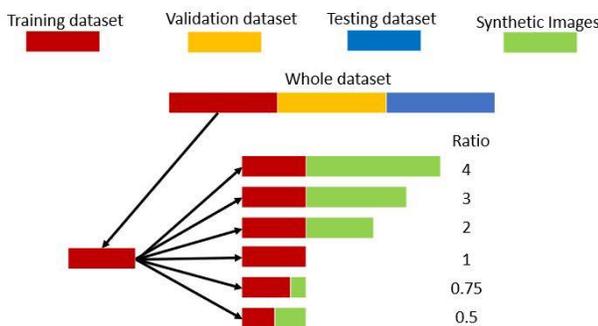


Fig. 4. Dataset Division.

Method 1: Baseline CNN

In baseline CNN method, only extracted symbols of P&IDs to be implemented. The standard CNN algorithm is

used to determine the symbol recognition rate against Method 2 and 3. The numbers of extracted P&ID symbols are calculated as ratio. TABLE II shows the ratio of validation dataset and testing dataset were fixed into 1 while the ratio of training dataset was keep reducing to validate the diversity accuracy of symbol recognition.

TABLE II. BASELINE CNN DATASETS

Baseline (Datasets)		
Training	Validation	Testing
N/A	1	1
N/A	1	1
N/A	1	1
1 (100% training dataset)	1	1
0.75 (75% training dataset)	1	1
0.5 (50% training dataset)	1	1

Method 2: CycleGAN + CNN

Same as Method 1, all extracted P&ID symbols split into training dataset, validation dataset, and testing dataset. However, this method involved synthetic images generated via CycleGAN model to enlarge variation symbols. The synthetic images are added into training dataset for increasing and equalizing the ratio. Then, CNN is used to recognize and classifying the symbols.

In this TABLE III, different ratio numbers of training dataset were tested to discover the difference accuracy recognition of P&IDs symbols.

TABLE III. CYCLEGAN+CYCLEGAN DATASETS

CycleGAN (Datasets)		
Training	Validation	Testing
4 (100% training dataset + 300% synthetic samples)	1	1
3 (100% training dataset + 200% synthetic samples)	1	1
2 (100% training dataset + 100% synthetic samples)	1	1
N/A	1	1
0.75 (75% training dataset)	1	1
0.5 (50% training dataset)	1	1

Method 3: Y.Zhang[9] algorithm

In Method 3, Y.Zhang[9] algorithm was implemented as for the validation against other methods. Since the P&ID symbols used are different from the origin paper, the pattern of the symbols was customized to tally these extracted symbols of P&IDs while the parameters of algorithm remained unchanged in [9]. The feature of this algorithm is to augment synthetic images with various scales, rotations, and random noises. These experiments were applied the ratio of training dataset same as Method 2 in TABLE IV.

TABLE IV. Y.ZHANG [9] ALGORITHM DATASETS

Y.Zhang [9] algorithm (Datasets)		
Training	Validation	Testing
4 (100% training dataset + 300% synthetic samples)	1	1
3 (100% training dataset + 200% synthetic samples)	1	1
2 (100% training dataset + 100% synthetic samples)	1	1
N/A	1	1
0.75 (75% training dataset)	1	1
0.5 (50% training dataset)	1	1

D. Proposed Methodology

Data augmentation is necessary because the limited symbols unable to provide the confidence during recognizing. We proposed to employ CycleGAN [23] with Convolution Neural Network(CNN).

In Fig. 5, the overview of proposed framework is displayed. CycleGAN is an unsupervised deep learning method which conducting a bidirectional translation between the two source domains which are domain X and target domain Y. Generally, the images collection from the source domain and target domain are not requirement that they are associated in any manner. The characteristics of CycleGAN implements two generator networks and a discriminator network for appraisal. Generator (G) and Discriminator (D) networks compete with one another. D is a classifier which trying to distinguish the samples between the synthetic images that generated via generator (fake) and the actual distribution (real). On the other hand, role of G attempts to fool the discriminator by producing synthetic output image. The input of the generator is along with source domain image x from Domain X and its output is a synthetic image. In additional, the inputs of a discriminator D are the synthetic output and an unpaired random image from the target image y of domain Y.

Even though CycleGAN allows unpaired random image but we apply a set of symbol images from a random class for consistent comparison.

Each generator in CycleGAN model involves an encoder, a transformer, and a decoder. Role of the generator ensures the features of images are extracted and converted into transformer (latent space). Transformer uses an attention mechanism to transfer the sequence to decoder. Then, a new feature vector of image is converted and reconstructed as output image in decoder.

In cycleGAN, the discriminator model is implemented as a PatchGAN model [23] which aims at classifying images as real or synthetic. The discrimination starts undergoing an appraisal with identifying the image belongs to real or synthetic. For discrimination, we apply adversarial losses [23] into both mapping functions to match up between the output generated images and real images from domain Y . For the mapping function $G: X \rightarrow Y$ and the discriminator Domain Y, where Generator(G) seeks to generate images $G(x)$ which looking closely images of domain Y, while discriminator D_Y is applied to distinguish between generated images of $G(x)$ and real images y of domain Y. G aims to minimize this training objective but D that tries to maximize it. Therefore, adversarial loss function is used to inverse mapping the first generator resulting $\min_G \max_{D_Y} L_{GAN}(G, D_Y, X, Y)$. A same adversarial loss for the mapping function on second generator (F). According to Figure 5, $F: Y \rightarrow X$ and its discriminator D_X as well resulting $\min_F \max_{D_X} L_{GAN}(F, D_X, X, Y)$.

Adversarial training learns mappings between G and F to generate the outputs closely to target domains Y and X respectively. Nevertheless, adversarial loss alone is uncertain to map the learned function onto an individual input x_i to a desired output y_i . but stimulate the tolerance of the cycle consistency. The cycle consistent argued the learned mapping functions to be further narrowed the space of the possible mapping functions for each image x of domain X. Thus, the image proceeds with translation cycle function to return image.

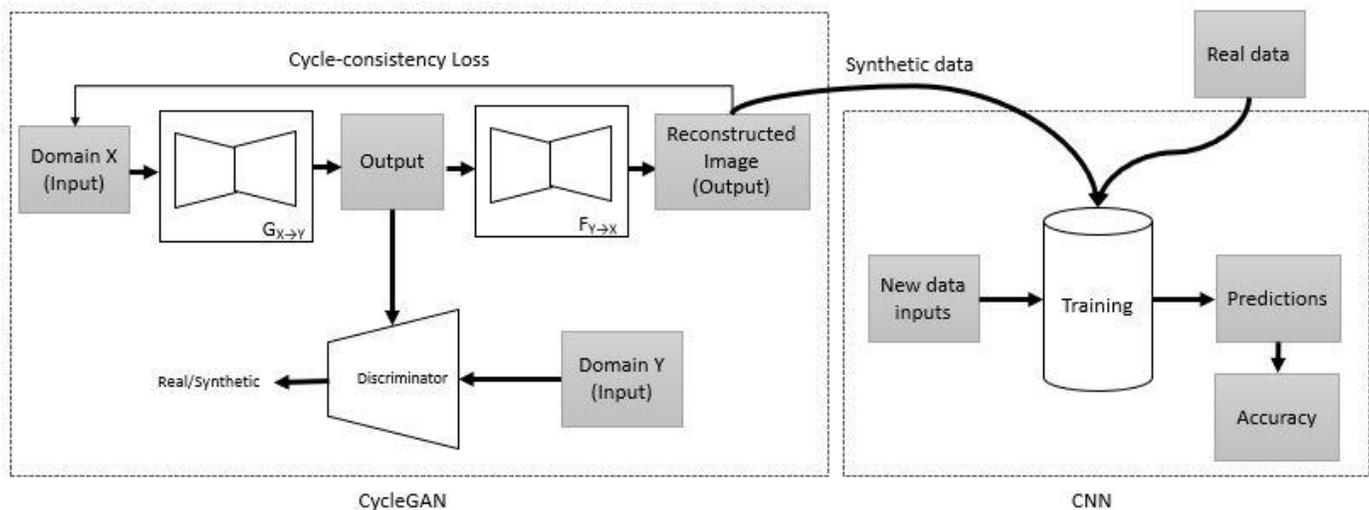


Fig. 5. Proposed Framework: CycleGAN (Generate Synthetic Samples) + CNN (Symbols Recognition).

x back to the original image intimately which resulting as $x \rightarrow G(x) \rightarrow F(G(x)) \approx x$. This referred to forward cycle consistency. Besides, every image y from domain Y , G and F should repeat with backward cycle consistency resulting $y \rightarrow F(y) \rightarrow G(F(y)) \approx y$. This function is impelled by a cycle consistency loss.

Algorithm 1: CycleGAN Pseudocode

```

1: procedure TRAINCYCLEGAN
2: Select samples  $\{x_i\}$  from domain  $X$ 
3: Select samples  $\{y_i\}$  from domain  $Y$ 
4: Compute generator  $G: X \rightarrow Y$ 
5: Compute the discriminator loss on Domain  $Y$ :

$$\mathcal{L}_{GAN}(G, D_Y, X, Y) = \mathbb{E}_{y \sim p_{data}(y)} [\log D_Y(y)]$$


$$+ \mathbb{E}_{x \sim p_{data}(x)} [\log(1 - D_Y(G(x)))] \quad (1)$$

6: Compute generator  $F: Y \rightarrow X$ 
7: Compute Cycle-consistency loss:

$$\mathcal{L}_{cyc}(G, F) = \mathbb{E}_{x \sim p_{data}(x)} [\|F(G(x)) - x\|_1]$$


$$+ \mathbb{E}_{y \sim p_{data}(y)} [\|G(F(y)) - y\|_1] \quad (2)$$


```

The generated synthetic samples via CycleGAN model to be transferred into second stage for CNN recognition predictions. In CNN stage, mixing synthetic samples and real training data to increase the numbers of dataset for training model. This adding more various patterns and finetune the training model to achieve more accurate performance.

E. CycleGAN + CNN Hyperparameters

The hyperparameter setups for CycleGAN [23] and CNN model were displayed in TABLE V and TABLE VI. These setups were standard in the Tensorflow implementation.

TABLE V. HYPERPARAMETERS FOR CYCLEGAN

Hyperparameters CycleGAN	
Learning rate	0.002
Epochs	50
Beta	0.5
Adversarial loss	L2 Distance
Cycle loss	10
Identity loss	5

TABLE VI. HYPERPARAMETERS FOR CNN

Hyperparameters CNN	
Learning rate	0.01
Batch size	32
Epochs	60

IV. EXPERIMENTS AND RESULTS

A. Evaluation Metrics

The evaluation metrics is implemented with the accuracy of recognition and confusion matrix.

According to the accuracy of recognition metric, Equation 3 is shown below:

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (3)$$

In this equation contains four important terminologies. TP represents the number of positives, TN represents the number of negatives, FP denotes the number of false positives, FN denotes the number of false negatives. The formula of Equation 3 is the total number of samples are divided by the sum of true positives and true negatives gives an overall assessment of the classifier performs across every classes.

Furthermore, another evaluation metric is confusion matrix, also called error matrix. It is a is a summarized table that is used to evaluate a classification model's performance. The number of positives and negatives predictions are totaled and broken down by class using count values. In confusion matrix table, predicted classifications are indicated by rows while true positives are indicated in every classes respectively by columns.

B. Experiments Results

The experiment results are displayed in TABLE VII. This table showed:

According to TABLE VIII, the results summarized the highest average accuracy is up to 92.85% which belonging to CycleGAN + CNN method. Compared to baseline CNN method with a ratio of 1:1:1, there are no synthesis images or symbol reduction in the dataset, the average accuracy is only 90.75%.

Furthermore, it is undeniable that the average accuracy is decreasing when reducing the number of training symbols in the Baseline CNN method. However, the CycleGAN+CNN method still has higher average accuracy at ratios of 0.5 and 0.75 as a training dataset compared to the Baseline and Y.Zhang [9] algorithm. In contrast, the CycleGAN model was able to generate a large number of synthetic images, but a ratio of 3 and a ratio of 4 average accuracy showed a significant drop in results as the training dataset size increased.

In Y.Zhang [9] method, the average accuracy in all ratios of training dataset are significantly lower compared to Baseline CNN and CycleGAN + CNN methods. These results showed that not all methods that can generate synthetic images improve accuracy due to several factors.

In Fig. 6, a clustered chart displayed the accuracy over the different ratio of training dataset based on the values from TABLE VIII. Various colors represent different methods. CycleGAN + CNN method achieved the peak accuracy values against another methods with all different ratios of training dataset.

TABLE VII. EXPERIMENTAL RESULTS - RECOGNITION ACCURACIES

Dataset Ratio			Baseline CNN (Without synthetic samples)					CycleGAN + CNN					Y.Zhang [9]				
Training	Validation	Testing	Test 1	Test 2	Test 3	Test 4	Test 5	Test 1	Test 2	Test 3	Test 4	Test 5	Test 1	Test 2	Test 3	Test 4	Test 5
4	1	1	N/A	N/A	N/A	N/A	N/A	90.98%	90.02%	90.66%	91.47%	89.21%	73.27%	72.79%	74.07%	75.04%	73.11%
3	1	1	N/A	N/A	N/A	N/A	N/A	91.30%	92.59%	91.79%	92.59%	91.14%	73.91%	73.91%	74.40%	73.59%	74.72%
2	1	1	N/A	N/A	N/A	N/A	N/A	92.92%	93.56%	92.75%	92.59%	92.43%	76.49%	75.52%	79.55%	77.62%	76.17%
1	1	1	91.46%	90.49%	90.98%	89.04%	91.78%	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
0.75	1	1	89.21%	88.72%	90.66%	90.98%	89.69%	89.69%	91.30%	90.66%	92.11%	90.98%	83.25%	86.63%	87.28%	84.22%	87.18%
0.5	1	1	88.56%	88.24%	87.27%	83.89%	87.60%	88.08%	87.92%	89.21%	86.47%	91.30%	77.30%	77.13%	73.43%	75.52%	76.49%

TABLE VIII. AVERAGE AND STANDARD DEVIATION OF RECOGNITION ACCURACIES FOR FIVE TIMES REPEATS

Dataset Ratio			Baseline CNN (Without synthetic samples)		CycleGAN + CNN		Y.Zhang [9]	
Training	Validation	Testing	Avg. Accuracy (%)	Avg. Std (%)	Avg. Accuracy (%)	Avg. Std (%)	Avg. Accuracy (%)	Avg. Std (%)
4	1	1	N/A	N/A	90.47%	0.88%	73.66%	0.91%
3	1	1	N/A	N/A	91.88%	0.69%	74.11%	0.45%
2	1	1	N/A	N/A	92.85%	0.44%	77.07%	1.58%
1	1	1	90.75%	1.07%	N/A	N/A	N/A	N/A
0.75	1	1	89.85%	1.10%	90.95%	0.89%	85.71%	1.85%
0.5	1	1	87.11%	1.87%	88.60%	1.80%	75.97%	1.58%

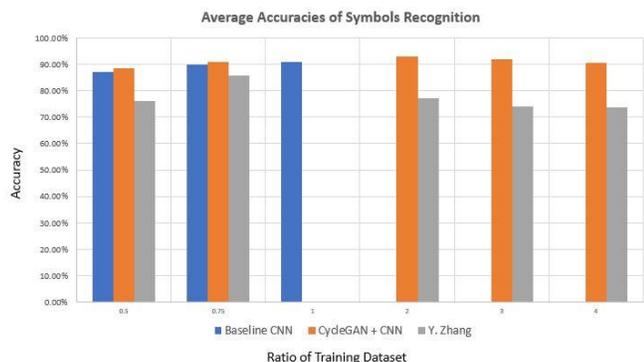


Fig. 6. Cluster Chart for Average Accuracies of Symbol Recognition.

The confusion matrix between Baseline method and CycleGAN method were shown in TABLE IX and TABLE X.

According to TABLE IX and TABLE X, each alphabet represents a type of symbol which is displayed in Section III. Both tables highlighted the difference of symbol recognition. As per comparison, (A) Check valve, (B) Gate valve, and (F) Concentric reducer showed the significant gaps. One of the reasons could be the variation issues. It requires more variation symbols to match with the P&ID symbols in testing dataset. In fact, CycleGAN model generate more variation synthetic images allow CNN architecture to recognize precisely especially these three types of symbols.

TABLE IX. CONFUSION MATRIX FOR THE BASELINE CNN MODEL

	A	B	C	D	E	F	G
A	19	1	0	0	0	4	11
B	0	137	13	0	0	0	3
C	0	0	96	1	1	0	0
D	0	0	0	111	0	0	0
E	0	0	0	0	76	0	0
F	1	0	0	0	0	55	30
G	3	0	0	0	0	0	59

TABLE X. CONFUSION MATRIX FOR THE CYCLEGAN + CNN MODEL

	A	B	C	D	E	F	G
A	23	0	0	0	0	3	9
B	0	146	5	0	0	0	2
C	0	0	97	0	0	0	0
D	0	1	0	110	0	0	0
E	0	0	0	0	76	0	0
F	1	0	0	0	0	70	16
G	2	0	0	0	0	1	59

V. DISCUSSION

According to Section IV, several experiments were tested on different methods. Our proposed framework resulted the combination of CycleGAN and CNN performed effectiveness on accuracy of recognition. Highest average accuracy up to 92.85% was achieved. The result showed CycleGAN the potential when trained on the synthetic samples. It performed better than the algorithm used in Y.Zhang [9]. Using CycleGAN model, generated synthetic samples increased the dataset for training. This provides CNN model better classifying and recognizing with sufficient training. Nevertheless, the confusion matrix of our proposed framework displayed a greater total number of true positives which recognized the specific symbols accurately.

Another interesting aspect from the results, different ratios of training dataset were conducted during experiments testing. Significantly, the accuracy rate of recognition gradually lower down when applying ratio 0.5:0.75 as the training data samples. The work proved the superior of synthetic samples from proposed method. Added CycleGAN synthetic samples to increase back to 100% or ratio 1 of the training dataset. It performed great contribution to accuracy rate even achieved higher than the standard 100% of training dataset from baseline CNN method without adding any synthetic samples. In addition, the incremental ratio of synthetic samples is scaled up to 400%. The best accuracy rate was performed with ratio 2 while ratio 3:4 start declined the accuracy of recognition.

VI. CONCLUSION AND FUTURE WORKS

In this study, the accuracy enhancement for recognizing symbols of engineering drawings required a data augmentation technique. We proposed CycleGAN + CNN method for synthetic symbols to enlarge the quantities of symbols in Piping & Instrument Diagrams (P&IDs). Our work addresses the lack of labelled data in deep networks by utilizing CycleGAN to generate synthetic images to supplement training data. We achieved the highest average accuracy of this method as high as 92.85%, which is a significant enhancement over CNN recognition alone. However, excessiveness or insufficient images can reduce the accuracy of recognition rate due to some factors such as overfitting and underfitting.

Future work is schemed to apply spatial transformations in CycleGAN architecture. If there are too many changes, it is difficult for the generator to learn the basic style efficiently. Additionally, spotting all types of symbols in an engineering drawing. An advanced object detection algorithm will be studied especially design for symbols. This will allow to detect symbols precisely regardless the pixel of engineering drawings and the size of symbols.

ACKNOWLEDGMENT

This research/paper was fully supported by Universiti Teknologi PETRONAS, under the Yayasan Universiti Teknologi PETRONAS (YUTP) Fundamental Research Grant Scheme (YUTP-FRG/015LC0-280).

REFERENCES

- [1] E. Elyan, L. Jamieson, and A. Ali-Gombe, "Deep learning for symbols detection and classification in engineering drawings," Elsevier, vol. 129, pp. 91-102, 1 June 2020.
- [2] C. F. E. Moreno-García, E.; Jayne, C., "New trends on digitisation of complex engineering drawings," *Neural Comput. Appl.*, vol. 31, pp. 1695-1712, 2019.
- [3] S. K., "Complex and composite graphical symbol recognition and retrieval: a quick review," in In: *International Conference on Recent Trends in Image Processing and Pattern Recognition (RTIP2R'16)*, 2016: Springer, Singapore, pp. pp 3-15.
- [4] W. L. Santosh K., "Graphical symbol recognition," *Wiley Encyclopedia of Electrical and Electronics Engineering*, 2015.
- [5] S. K., "Document image analysis," in In: *Document Image Analysis*, 2018: Springer, Singapore, pp. pp 1-15.
- [6] E. Elyan, Garia, C.M., Jayne,C., "Symbols Classification in Engineering Drawings," presented at the International joint conference on neural networks 2018 (IJCNN), Rio de Janeiro, Brazil, 8-13 July 2018, 2018.
- [7] A. Rezvaniyar, M. Cote, and A. Branzan Albu, "Symbol spotting for architectural drawings: state-of-the-art and new industry-driven developments," *IPSN Transactions on Computer Vision and Applications*, vol. 11, no. 1, p. 2, 2019/05/10 2019, doi: 10.1186/s41074-019-0055-1.
- [8] E. Yu, Cha, J.M., Lee, T., Kim, J., Mun, D., "Features Recognition from Piping and Instrumentation Diagrams in Image Format Using a Deep Learning Network " 2019.
- [9] Y. Zhang, "CNN-based Symbol Recognition and Detection in Piping Drawings," 16-Aug-2019.
- [10] D.-Y. Yun, S.-K. Seo, U. Zahid, and C.-J. Lee, "Deep Neural Network for Automatic Image Recognition of Engineering Diagrams," *Applied Sciences*, vol. 10, no. 11, p. 4005, 2020. [Online]. Available: <https://www.mdpi.com/2076-3417/10/11/4005>.
- [11] H. Kim et al., "Deep-learning-based recognition of symbols and texts at an industrially applicable level from images of high-density piping and instrumentation diagrams," *Expert Syst. Appl.*, vol. 183, p. 115337, 2021.
- [12] M. Delalandre, É. Trupin, and J.-M. Ogier, "Symbols Recognition System for Graphic Documents Combining Global Structural Approaches and Using a XML Representation of Data," Berlin, Heidelberg, 2004: Springer Berlin Heidelberg, in *Structural, Syntactic, and Statistical Pattern Recognition*, pp. 425-433.
- [13] W. Fang, F. Zhang, V.-S. Sheng, and Y. Ding, "A Method for Improving CNN-Based Image Recognition Using DCGAN," *Computers, Materials & Continua*, vol. 57, no. 1, pp. 167-178, 2018. [Online]. Available: <http://www.techscience.com/cmcc/v57n1/22963>.
- [14] N. Shuping, "Feature Extraction and Segmentation Processing of Images Based on Convolutional Neural Networks," *Optical Memory and Neural Networks*, vol. 30, no. 1, pp. 67-73, 2021/01/01 2021, doi: 10.3103/S1060992X21010069.
- [15] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, Warde-Farley, and D. e. al., "Generative adversarial nets.," *Advances in Neural Information Processing Systems*, pp. 2672-2680, 2014.
- [16] A. Shrivastava, T. Pfister, O. Tuzel, J. Susskind, W. Wang, and R. Webb, "Learning from simulated and unsupervised images through adversarial training," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2107-2116, 2017.
- [17] M. Frid-Adar, I. Diamant, E. Klang, M. Amitai, J. Goldberger, and H. Greenspan, "GAN-based synthetic medical image augmentation for increased CNN performance in liver lesion classification," *Neurocomputing*, vol. 321, pp. 321-331, 2018.
- [18] C. Han et al., "Learning more with less: Conditional PGGAN-based data augmentation for brain metastases detection using highly-rough annotation on MR images," *arXiv preprint arXiv:1902.09856*, 2019.
- [19] J. Liu, J. Li, T. Liu, and J. Tam, "Graded Image Generation Using Stratified CycleGAN," *Cham*, 2020: Springer International Publishing, in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020*, pp. 760-769.

- [20] X. Zhang, X. Han, C. Li, X. Tang, H. Zhou, and L. Jiao, "Aerial Image Road Extraction Based on an Improved Generative Adversarial Network," *Remote Sensing*, vol. 11, no. 8, p. 930, 2019. [Online]. Available: <https://www.mdpi.com/2072-4292/11/8/930>.
- [21] M. Park, D. Q. Tran, D. Jung, and S. Park, "Wildfire-Detection Method Using DenseNet and CycleGAN Data Augmentation-Based Remote Camera Imagery," *Remote Sensing*, vol. 12, no. 22, p. 3715, 2020. [Online]. Available: <https://www.mdpi.com/2072-4292/12/22/3715>.
- [22] S. Liu, H. Wu, S. Luo, and Z. Sun, "Stable Video Style Transfer Based on Partial Convolution with Depth-Aware Supervision," *Proceedings of the 28th ACM International Conference on Multimedia*, 2020.
- [23] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," In *IEEE International Conference on Computer Vision (ICCV)*, 2017.

A Graph-oriented Framework for Online Analytical Processing

Abdelhak KHALIL¹

LEFCG-SIAD Laboratory
Hassan First University of Settat
Settat, Morocco

Mustapha BELAISSAOUI²

LEFCG-SIAD Laboratory
Hassan First University of Settat
Settat, Morocco

Abstract—OLAP (Online Analytical Processing) is a tried-and-tested technology and a core concept in Business Intelligence. With data flowing from different and countless sources, exploring data in order to deliver actionable insights has become a daunting task with current OLAP tools despite the cycle of improvement that has gone through it. In the last decade, with the emergence of the big data phenomenon, NoSQL databases are seeing a spike in popularity and become more used in industry and academia as their value in handling a huge and varied amount of data become increasingly evident. Graph oriented database is one of the four chief types of NoSQL oriented databases that represent a promising technology candidate for big data analytics. In this paper we bring forward our contribution to graph-oriented analytical processing, which is twofold. First, we provide a novel approach for modeling a graph-oriented data warehouse. Second, we propose a data cube materialization through the precomputation of aggregated nodes. We present how typical OLAP queries can be performed against data warehouses stored in NoSQL graph-oriented database management systems. An implementation is conducted on a fictional data warehouse using Neo4j and the Cypher declarative language. The same dataset is stored in a relational data warehouse in order to compare storage space and query performance. Thus, the obtained results shows that graph OLAP implementation outperform clearly the relational alternative in term of query response time.

Keywords—Graph OLAP; data warehousing; graph databases; NoSQL; data cube; decision support system

I. INTRODUCTION

OLAP stands for (Online Analytical Processing) and describe a software technology dedicated to decision-making purpose. It is designed to locate meaningful intersections between multiple axes of analysis. The dimensional modelling is an integral part of OLAP systems and defines at the conceptual level the fact concept which holds measurements or metrics regarding a business process event, and the dimension concept which provides a context describing the fact. Data conversion from an OLTP (Online Transaction Processing) database of two-dimensional to the multi-dimensional model is done by an ETL (Extract, Transform, Load) tool. OLAP servers have historically been implemented mainly using four approaches: Relational-OLAP(ROLAP), Multidimensional-OLAP(MOLAP), Hybrid-OLAP(HOLAP) and Desktop-OLAP(DOLAP) [1], [2]. Each implementation has its strengths and its limitation and must be evaluated based on the business requirements.

With the IT revolution, and being aware of the potential of information, organizations around the globe has moved from the archaic age, which relies on industrial economy into a new era characterized by data driven economy. This race after technology in order to gain competitive advantages has contributed to the generation of large volumes of data. As a consequence, data analytics are becoming a huge challenge for traditional OLAP systems due its vertical scalability and its low computation ability. Indeed, earlier-generation of OLAP implementations are of poor storage and computational capacities, because they are built upon on old architectures and cannot match the requirement of big data analytics, especially data storage and data retrieval requirements. Another common problem is OLAP cube building over big data which could reach a critical complexity due to the increasing number of dimensions and the unstructured nature which characterize big data sets [3],[4].

To overcome the challenges of scale and complexity associated with today's data, OLAP researches moved in a new direction. Namely, the use of NoSQL databases in OLAP solutions which is considered as a promising alternative for traditional data storage tools [5]–[8], [9]. This revolutionary technology offers several interesting features that cannot be achieved with classical database management systems like cluster computing and the ability to process both semi-structured and unstructured data. In this paper, we are focused particularly in graph database, a class of NoSQL databases that uses a graph model composed of nodes and edges instead of relational model [10][11], and we claim that the graph data structure is suitable for data warehousing and online analysis.

Implementing an OLAP cube using a graph database is not a straightforward process. The multidimensional model used to instantiate the data cube must be converted to a logical model suitable to graph oriented database. Furthermore, typical OLAP queries must be translated to a specific language supported by this technology. The aim of this work is to illustrate the potentiality of graph databases to handle OLAP structures designed for reporting. In this context, we define a set of mapping rules in order to migrate dimensionally modelled data into the graph database. And we demonstrate how typical OLAP operations can be performed against a graph database. In Fig. 1, we position our proposal regarding the literature. The key contributions of this work can be summarized as follows:

- We propose an implementation of OLAP engines under graph database using two different logical models that are equivalent to ROLAP and MOLAP models. We define a set of rules used for the mapping from the multidimensional model to these models. An experiment is conducted to highlight the differences between the two meta-models using a case study.
- We propose an effective aggregation technique to build the lattice of cuboids from a data warehouse built upon a graph database management system.
- Then, we provide an extension of the declarative Cypher language to basic OLAP queries. We consider in this work Neo4j as a graph database engine.

The remainder of this paper is structured as follows. In the next section we present the background of our work, and we provide an overview of the state of the art related on Graph-OLAP. In Section III we present our modeling approach for graph OLAP. In Section IV we give an implementation of the proposed approach using the Cypher language. In Section V, we discuss experimental results. The last section concludes this work and suggests eventual research directions.

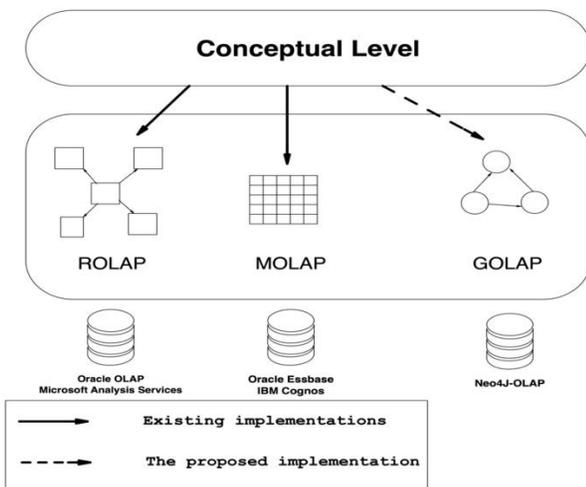


Fig. 1. Convection from the Conceptual Level to different OLAP Implementations.

II. RELATED WORK

A. The Multidimensional Schema

The multidimensional schema is the starting point to design and implement data warehouse systems. It defines four major concepts: fact, measures, dimensions and hierarchies [12].

Formally, a Multidimensional Schema denoted S is a triplet $(F^S, D^S, Star^S)$ where:

- $F^S = \{F_1, \dots, F_n\}$ a finite set of facts.
- $D^S = \{D_1, \dots, D_m\}$ a finite set of dimensions.

$Star^S : F_i \rightarrow 2^{D_i}$ is an incidence function mapping each fact $F_i \in F^S$ to its associated dimensions $D_j \in D^S$.

A **fact** is the business process studied and is represented by a pair (N^{F_i}, M^{F_i}) where:

- N^{F_i} is the name of the fact.
- $M^{F_i} = \{m_1^{F_i}, \dots, m_n^{F_i}\}$ a finite set of measures.

A dimension $D_i \in D^S$ is defined by $(N^{D_i}, Att^{D_i}, H^{D_i})$ where:

- N^{D_i} is the name of the dimension.
- $Att^{D_i} = \{a_1^{D_i}, \dots, a_m^{D_i}\}$ a finite set of attributes.
- $H^{D_i} = \{l_1^{D_i}, \dots, l_k^{D_i}\}$ a set of hierarchy levels.

A hierarchy organizes measures at different level of aggregations. A hierarchy level $l_j^{D_i} \in H^{D_i}$ can be defined by $(N^{l_j^{D_i}}, Att^{l_j^{D_i}}, Weak^{l_j^{D_i}})$ where:

- $N^{l_j^{D_i}}$ is the name of the hierarchy level.
- $Att^{l_j^{D_i}} = \{a_1^{l_j^{D_i}}, \dots, a_m^{l_j^{D_i}}\}$ an ordered set of attributes.
- $Weak^{l_j^{D_i}} : a_j^{l_j^{D_i}} \rightarrow \{wa_1, \dots, wa_k\}$ is a function possibly associating parameters to a set of weak attributes.

B. The Graph Model

NoSQL graph-oriented database are based upon the concepts of graph model which organize data into collections of nodes and edges. Once data loaded, graph theory algorithms make it easy to handle semantic queries by calculating the shortest path between nodes. Graph database specify connections at insert time and avoid by then the problem of join index lookup performance as querying data becomes a matter of graph traversal. This makes graph engines optimum when the meta-model of data being stored has many overlapping relationships. This contrast with relational database which store the links between tables at the logical level and relies on relational algebra operations to manipulate the data stored in the database management systems in a relevant logical format.

Formally, a graph database denoted G is a set of properties $(N, E, \phi, L_N, L_E, P_N, P_E)$ comprising:

- N a set of nodes (also called vertices).
- $E \subseteq N \times N$ a set of edges (also called links).
- $\phi : E \rightarrow \{\{x, y\} \mid x, y \in N, x \neq y\}$ a function linking an edge to a pair of nodes.
- $L_N = \{l_1, \dots, l_n\}$ a set of node labels.

- $L_E = \{l_1, \dots, l_m\}$ a set of edge labels.
- $P_N = \{p_1^N, \dots, p_j^N\}$ a set of node properties.
- $P_E = \{p_1^E, \dots, p_k^E\}$ a set of edge properties.

A node $n_i \in N$ is a pair (l_i, a_{n_i}) , where $l_i \in L_N$ is the node label, and $a_{n_i} = [a_1, \dots, a_n]$ a set of attributes associated with the node. Identically an edge $e_j \in E$ is represented as $(l_j, a_{e_j}, \eta_x, \eta_y)$, where $l_j \in L_E$ the edge label, a_{e_j} a set of edge attributes, η_x the starting node and η_y the ending node.

C. Graph-Based OLAP

Over the last few years, big data analytics have known a meteoric adoption of NoSQL. Considerable attempts to model an OLAP cube with this technology have appeared. Several research works have been conducted to implement OLAP systems using columnar databases [5],[6],[7], others using the document-oriented database [8],[13],[14],[15],[16] and last but not least key-value stores [17],[18],[19].

Although graph databases are widely used in OLTP systems, especially when the need of modeling multiple connections is self-evident, it does not exist, to the best of our knowledge, any OLAP solution which uses a graph database at the physical level in the market. However, graph OLAP concept has been around for years. Indeed, some interesting works attempted to implement OLAP systems using graph technology. A decade ago, Chen et al.[20], [21] studied the possibility to perform multi-dimensional analysis on graph data, the authors developed a graph OLAP framework having two major subcases: Informational OLAP and Typological OLAP and proposed the basic definition of OLAP operations under this framework.

Many recent research works have been interested in implementing OLAP engines under property graph databases. In [22], the authors introduce a new data warehousing concept called Graph Cube which stands for an OLAP infrastructure that support analytical queries over a multidimensional network. In [23], the authors define the concept of GOLAP which is an extension of Online Analytic Processing(OLAP) under graph database, some features are listed such as semantics queries and structural analytics. In this work the authors address the challenges of speed and storage related to GOLAP and proposes possible solution to deal with them like graph data reduction and query result approximation when the execution time is too long, unfortunately the authors did not provide an implementation of the proposed framework and focus rather on the possible formalization. In [24], the authors propose a novel graph cube framework called Two-Step Multi-dimensional Heterogeneous(TSMH) which consists of an Entity Hyper Cube and Dimension Cube. In the Entity Hyper Cube n-meta path relation algorithm is used to guide the aggregation of the network and to extend drill-down/roll-up operations. In the Dimension Cube the efficiency of dimension operation is improved by using a hierarchical coding for entity type and dimensions.

Along the same vein, in [25] the author proposed an OLAP data structure that relies on typed nodes to store facts and dimensions, and introduced an extension of the Cypher language to basic OLAP queries. The authors didn't provide any experimental campaign to validate their proposal as they rather focused on the demonstration of its feasibility. In [26], [27], the authors proposed a formal multidimensional data model for graph analysis based on node and edge-labeled called graphoids, and presented a proof of concept implementation using a Neo4j graph database.

Regarding the instantiation of data warehouses using property graph database, in [28] the authors define a set of transformation rules for mapping between the multidimensional conceptual model and NoSQL graph model.

All the cited works present an interesting background for graph-based online analytical processing. The majority of them addressed the issue of the adaptation of graph structure to OLAP needs. Although they share some similarities with ours, the contribution of this work is quite different as we propose a novel approach for implementing both a data warehouse and OLAP engine based on efficient data cube materialization over graph database.

III. GRAPH OLAP MODEL

OLAP engines have been traditionally categorized whether they perform pre-computation of OLAP cuboids or not. Following this taxonomy, OLAP systems where all part of the cube is pre-computed and stored in memory or disk are called multidimensional OLAP systems (MOLAP) and systems where computation of OLAP cuboids is performed on-demand directly from the data warehouse are considered as Relational OLAP models (ROLAP).

In this section we define the logical graph model for data warehousing. We consider two approaches by analogy to the ROLAP and MOLAP models; each one differs in term of structure and content when the mapping from the conceptual model is performed. In the first approach, fact, dimensions and the link between them are materialized by nodes and edges following several mapping rules, while in the second approach we talk rather about an aggregate lattice modeled using the graph paradigm. In what follows, we will use a fictional electronics company as a running example. The star schema of our cube is depicted in Fig. 2:

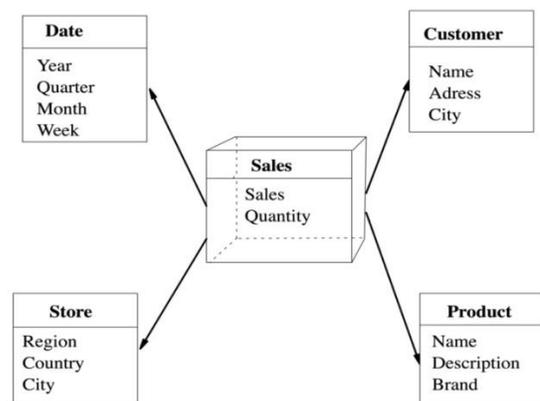


Fig. 2. The Star Schema.

A. First Approach

This approach corresponds to the lightly summarized data model. It defines a meta-model in which each component (fact and its associated dimensions) will be transformed to a node. The relation between nodes will be materialized by edges as detailed by the following mapping rules:

Rule.1. Each fact component $F_i \in F^S$ is converted to a node defined by (l_i, a_{n_i}) where:

- l_i is the name of the fact.
- Each measure $m_k^{F_i} \in M^{F_i}$ is converted to a node attribute $a_k \in a_{n_i}$.

Rule.2. Each dimension component $D_j \in D^S$ is translated to a node defined by (l_j, a_{n_j}) where:

- Each dimension attribute $a_k^{D_j} \in Att^{D_j}$ is mapped into a node attribute $a_k \in a_{n_j}$.
- Each hierarchy level $l_k^{D_j} \in H^{D_j}$ will be stored as a node alike dimension.
- Hierarchy levels are connected by edges to express how they are hierarchically linked.

Rule.3. The link between fact and its associated dimensions is represented by an edge (l_i, η_x, η_y) where:

- $l_j \in L_E$ is the name of the relation.
- η_x a node representing the fact.
- η_y a node representing an associated dimension

For the star schema represented in Fig. 2, the application of the aforementioned rules will give us the following meta-model, Fig. 3:

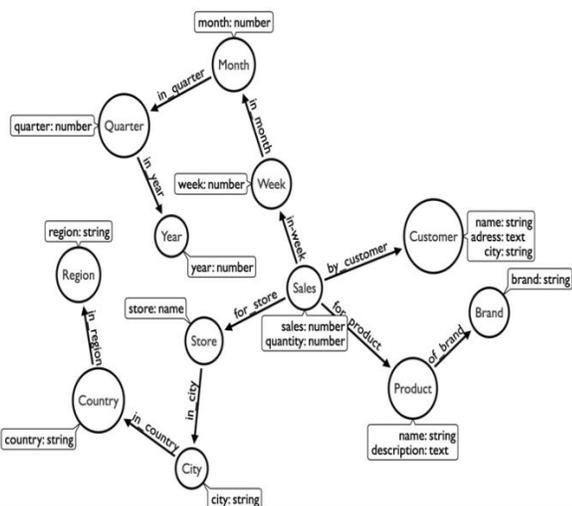


Fig. 3. The Graph-OLAP Schema According to the First Approach.

B. Second Approach

When we want to perform aggregation on a graph OLAP built according to the first approach, the query we should write is served on-demand and relies on fact nodes which are retrieved then aggregated using an aggregation function. This technique achieves the required result, but it is not optimized for a large data volume. Moreover, it is tending to the opposite of OLAP philosophy where data aggregation is pre-computed and stored.

The second approach corresponds to a highly summarized data model where measure aggregations are pre-calculated and directly available for the sake of query performance. The set of pre-computed aggregations is called an *aggregate lattice*. Concretely, fact measures are aggregated according to different combinations of dimensions and stored as a node with two labels.

- l_i identify the multidimensional concept $l_i = 'Aggregate'$.
- l_j a label which follows a particular pattern that identify uniquely which cuboid the aggregate is calculated for. This label is in the form of a bitmask starting with a letter that indicate the type of the aggregate (S for Sum, A for Average, etc.). The remaining part is an ordered sequence of n position (one of each hierarchy level), each position can have three possible values: (x) if the aggregate is calculated for all occurrences of the level, (1) if the aggregate is performed for each occurrence of the level and (0) if the aggregate is not calculated for the level.

If we refer to our running example and considering only high levels of granularity. Let's assume by convention that the order of position levels is:

Product.Brand-Product.Product-Store.Region-Store.Country-Date.Year-Date.Quarter.

An example of bitmask construction is depicted in Table I and Fig. 4 displays such a representation:

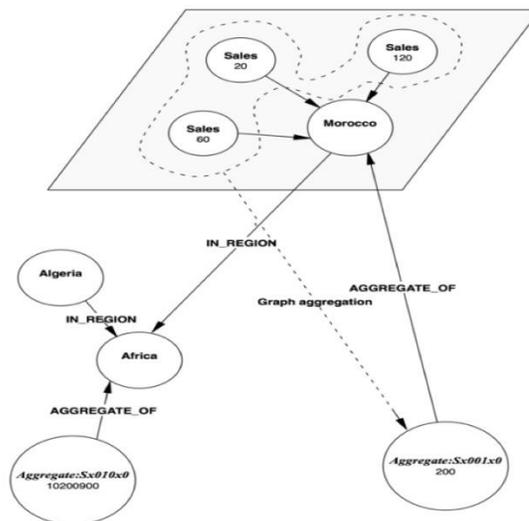


Fig. 4. Graph Aggregation According to the Second Approach.

TABLE I. CUBOID BITMASK CONSTRUCTION

Node label	Description	Scope
Aggregate:S10x0x0	Sum by Product.Brand	One level aggregate
Aggregate:Sx010x0	Sum by Store.Region	One level aggregate
Aggregate:Ax0x010	Avg by Date.Year	One level aggregate
Aggregate:S1010x0	Sum by Product.Brand and Store.Region	Two levels aggregate

IV. IMPLEMENTATION

A. Answering Typical Analytical Operation using Cypher

OLAP operations help users to view data from different perspectives providing a convenient environment for real-time data visualization and analysis. OLAP defines several basic operations; the most popular ones are roll-up, dicing and slicing. In this section we present how these operators can be expressed over a data cube designed according to the first approach.

Queries are written using the Cypher syntax, a declarative query language intended to be executed on a database engine built on the graph model. Cypher relies on the concept of pattern matching for querying and updating graphs [12]. A detailed description of the Cypher syntax is beyond the scope of this paper.

1) *Roll-up*: The roll-up operation (also called consolidation or aggregation operation) performs aggregation on a data cube in two ways, either by reducing the number of dimensions or by climbing up a concept hierarchy for a dimension. It is like zooming-out feature from the most detailed granularity level to the less detailed one.

In the query given in Listing.1, the rollup operation is performed by climbing up the concept hierarchy of *Product* dimension (*Product* → *Brand*), and of *Store* dimension (*Store* → *City*). The execution of the query results in the creation of a node containing the aggregated measures and two new relations linking the created node with its associated dimension hierarchies.

Listing. 1. Roll up-Aggregation of sales and quantities by product brand and store city.

1. MATCH (br:Brand)-[:]-(:prod:Product)-[:]-(:fact:Sales)
2. MATCH (ct:City)-[:]-(:st:Store)-[:]-(:fact:Sales)
3. WITH DISTINCT br, ct, SUM(fact.sales) AS SumSales, SUM(fact.quantity) AS SumQuantity
4. CREATE (br)-[:AGGREGATE_OF]-(:agg:Aggregate:S1010x0 {sales: SumSales, quantity: SumQuantity})-[:AGGREGATE_OF]->(ct)
5. RETURN br,agg,ct;

2) *Dicing*: Dicing is the operation of selecting a subset over all the dimensions and picking only specific dimension parameter values. We can think of dicing as zoom feature using smaller scale.

In Listing.2 the dice operation is performed using a selection criterion over *Brand* and *Year* dimensions. The generated cube has two dimensions.

Listing. 2. Dice-Selecting the sum of sales for the brand Apple in 2018.

1. MATCH (br:Brand {brand: 'Apple'})<[*]-(:fact:Sales)
2. MATCH (year:Year {year: 2018})<[*]-(:fact:Sales)
3. RETURN SUM(fact.sales) AS Sales, SUM(fact.quantity) AS Quantity;

3) *Slicing*: Slicing is similar to dicing with a little difference. It emphasizes one specific dimension and provides a new sub-cube by filtering on a particular attribute. It can be considered as a specialized filter for specific dimension parameter value.

In Listing.3 Slice is carried out for the dimension *Region* using the criterion *Region*= 'Asia'.

Listing. 3. Slice- Selecting the sum of sales in region Asia

1. MATCH (reg:Region) <[*]-(:fact:Sales)
2. WHERE reg.name='Asia'
3. RETURN reg.region AS Region, SUM(meas.sales) AS Sales, SUM(meas.units) AS Units;

B. Aggregates Creation

We refer to the property graph in Fig. 3 and the set of aggregates in Table I, and then we show how we can perform pre-calculation of our sample cuboids.

1) *Aggregate by product brand*: Query results in cypher are evaluated by its core concept, namely, pattern matching. By using patterns, you describe the requested data shape, then the Cypher engine is responsible for restoring the data you are looking for. For example, to build the aggregate value *Aggregate:S10x0x0*, a join is implemented by means of matching *Sales* → *Brand* against the OLAP-graph. It is worth noting that the edge label linking the fact and the dimension nodes is not required as it is inferred from node types.

In SQL, this is equivalent to a join between the fact table *Sales* and the dimension table *Brand* followed by the aggregation function SUM and GROUP By clause over *Brand* attributes.

Listing. 4. Creation of the aggregate *Aggregate:S10x0x0*.

1. MATCH (brand:Brand)-[*2]-(:s:Sales)
2. WITH DISTINCT brand, SUM(s.sales) AS SumSales, SUM(s.quantity) AS SumQuantity
3. CREATE (a:Aggregate:S10x0x0 {sales: SumSales, quantity: SumQuantity})-[:AGGREGATE_OF]->(brand);

Fig. 5 shows how the aggregate *Aggregate:S10x0x0* (By product brand) fits in the property graph (colored in grey). It is a one-level aggregate as it is calculated against one hierarchical level(colored in red).

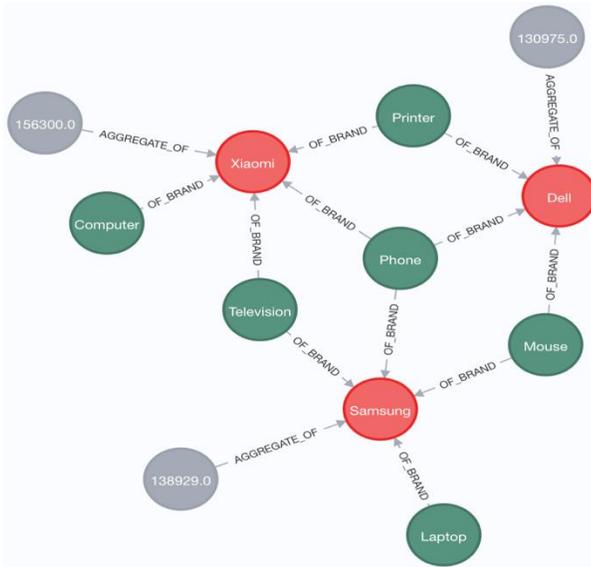


Fig. 5. Graph Visualization for Aggregates of Product Brands

2) *Aggregate by region*: In Listing. 5, the aggregate node *Aggregate:Sx010x0* (by region) is created:

Listing. 5. Creation of the aggregate *Aggregate:Sx010x0*

```

1. MATCH (r:Region)-[*4]-(s:Sales)
2. WITH DISTINCT r, SUM(s.sales) AS SumSales, SUM(s.quantity) AS
   SumQuantity
3. CREATE (a:Aggregate:Sx010x0 {sales: SumSales, quantity: SumQuantity})-
   [:AGGREGATE_OF]->(r);

```

3) *Aggregate by year*

Listing. 6. Creation of the aggregate *Aggregate:Sx010x0*

```

1. MATCH (y:Year)-[*3]-(s:Sales)
2. WITH DISTINCT y, AVG(s.sales) AS AvgSales, AVG(s.quantity) AS
   AvgQuantity
3. CREATE (a:Aggregate:Ax0x010 {sales: AvgSales, quantity: AvgQuantity})-
   [:AGGREGATE_OF]->(y);

```

4) *Aggregate by product brand and region*: In Listing.7, the two-levels aggregate node *Aggregate:S1010x0* (by product brand and region) is created:

Listing. 7. Creation of the aggregate *Aggregate:S1010x0*

```

1. MATCH (brand:Brand)-[*2]-(s:Sales)
2. MATCH (r:Region)-[*4]-(s:Sales)
3. WITH DISTINCT brand, r, SUM(s.sales) AS SumSales, SUM(s.quantity) AS
   SumQuantity
4. CREATE (brand)-[:AGGREGATE_OF]-(a:Aggregate:S1010x0 {sales:
   SumSales, quantity: SumQuantity})-[:AGGREGATE_OF]->(r);

```

Increasing the materialization of the aggregates can improve considerably query performance, but can also affect drastically storage space since aggregate nodes are stored on disk. The precalculation of all possible aggregate values is often not needed. Generally, OLAP engines chose the percentage of precomputed values based on business needs, the remaining aggregates are calculated in response to a query. We can imagine a scenario in which potentially requested aggregates are inferred from log files that contains previously executed queries.

V. RESULTS AND DISCUSSION

We conducted experiments to evaluate two aspects for the OLAP implementation under graph database: storage space and query performance. For this, the solution we propose is compared with a ROLAP implementation under Oracle relational database containing the same dataset. The experiment is carried out on a Unix machine (macOS) having a core-i7 CPU,16GB of RAM and 1 TB of stockage memory and running Neo4j community edition v4.3.

A. Data Generation

The dataset used in the experiment is generated using a novel NoSQL star schema benchmark named KoalaBench [29], [30], [30]. This tool is developed with Java language and is derived from the reference benchmark TCP-H. For clarity and to fit the meta-model in our running example the Supplier is replaced with the Store dimension, LineItem is renamed with Sales, and for the equivalent graph model, only few dimension parameters are tracked. Datasets can be generated in different configurations (different file format including tab, csv, json, xml..., and multiple models). The size of the generated data by scale factor is detailed in Table II.

TABLE II. SIZE OF THE DATA GENERATED BY SCALE FACTOR (SOURCE¹)

		Lines	Disk Space in Byte (SF=1)	Avg. Disk space/line (Byte)
Tables	Sales (LineItem)	SF _x 6000000	862558617,6	143,76
	Product (Part)	SF _x 200000	28521267,2	142,6
	Customer	SF _x 150000	16043212,8	1069,54
	Store (Supplier)	SF _x 10000	1677721,6	167,77
	Nation	25	367	14,68
	Region	5	73,4	14,68
	Date	SF _x 2556	168522	65,93
Size on disk			0,85 GB	-

B. Experiment 1: Memory Consumption Per Scale Factor

In this experiment we use a global flat CSV file representing data in a flat meta-model. In the appendix (Listing.8), we attach the Cypher script for loading data from an CSV file into Neo4j database according to our modeling approach. A fragment of the generated graph is represented in Fig. 6. The number of nodes and edges for the corresponding graph is depicted in Table III.

¹ http://www.tpc.org/tpc_documents_current_versions/pdf/tpc-h_v2.17.1.pdf

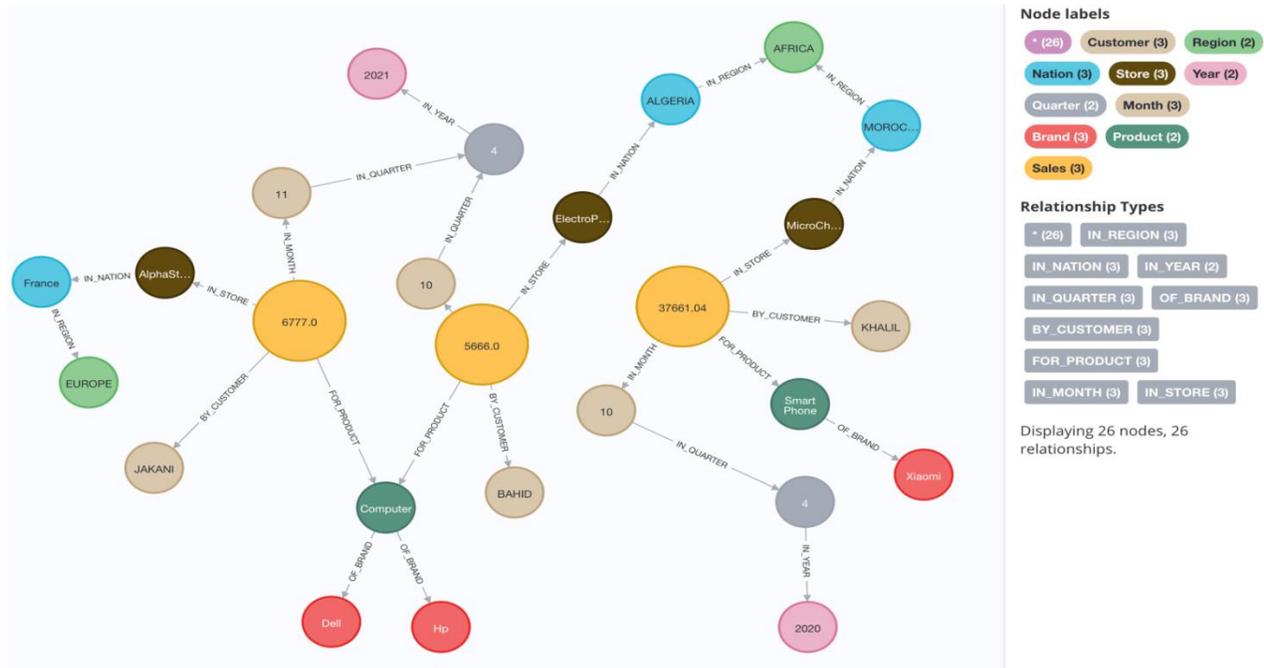


Fig. 6. A Portion of the Graph-OLAP.

TABLE III. MEMORY USAGE FOR THE GRAPH MODEL ON DIFFERENT SCALE FACTORS

		<i>Sf=1</i>	<i>Sf=5</i>	<i>Sf=10</i>
Nodes	Sales	6000000	30000000	60000000
	Product	200000	1000000	2000000
	Brand	100	500	1000
	Customer	150000	750000	1500000
	Store	10000	50000	100000
	Nation	25	25	25
	Region	5	5	5
	Month	79	395	790
	Quarter	27	135	270
	Year	7	35	70
Edges	FOR_PRODUCT	6000000	30000000	60000000
	OF_BRAND	200000	1000000	2000000
	BY_CUSTOMER	6000000	30000000	60000000
	IN_STORE	6000000	30000000	60000000
	IN_NATION	10000	50000	100000
	IN_REGION	25	25	25
	IN_MONTH	6000000	30000000	60000000
	IN_QUARTER	79	395	790
	IN_YEAR	27	135	270
Size on disk	3,3 GB	16.6 GB	33.2 GB	

From the Table III, we can see that a snowflake schema on a graph database requires more storage space than in a relational one (more than 3 times for SF=1). This is easily

explained: property graph databases store relationships physically on disk using edges while the concept of foreign key is used instead by relational databases. Furthermore the metadata is stored individually for each record in graph database unlike relational model which define the structure of the data at a higher level(the table itself). Which means that property names are repeated for each item. Indeed, graph databases are very storage intensive. This is traded for higher query performance. Since nowadays hard disks are inexpensive, it would be worthwhile trade-off to buy more storage space than keeping users waiting.

C. Experiment 2: Query Performance

The purpose of this experiment is to measure empirically the performance of graph-OLAP to process analytical queries when scaling up in comparison with the ROLAP implementation under Oracle database. We have exposed the system to a scale factor equal to 10 wich generates 11,6 Go of random data in csv file format. Query configuration includes queries involving gradually an increasing number of dimensions as depicted in Table IV. Each query was executed three times and the average of the elapsed time is presented in Fig. 7.

Experiment results show that the relational implementation defeats the Graph alternative when the query involves one dimension, but when the query dimensionality increases the graph alternative show better performance ranging from 1,82 to 2,29 times faster. Indeed, in relational databases the deeper we go in joining tables the more queries show slower processing time because it requires scanning of all table involved in the query which has a considerable cost. Unlike relational databases which suffer the pain of joining tables, graph databases express relationship at the physical level. That means, the links between nodes exists physically on disk and are named and directed which, makes graph traversal easier.

TABLE IV. QUERY CONFIGURATION

Query	Dimensionality	Dimension attributes	Measure
Q1	1D	Date:year	Sum(sales)
Q2	2D	Product:name Store:region	Sum(sales)
Q3	3D	Product:name Store:region Date:month	Sum(sales)
Q4	4D	Product:name Store:region Date:quarter Customer:name	Sum(sales)

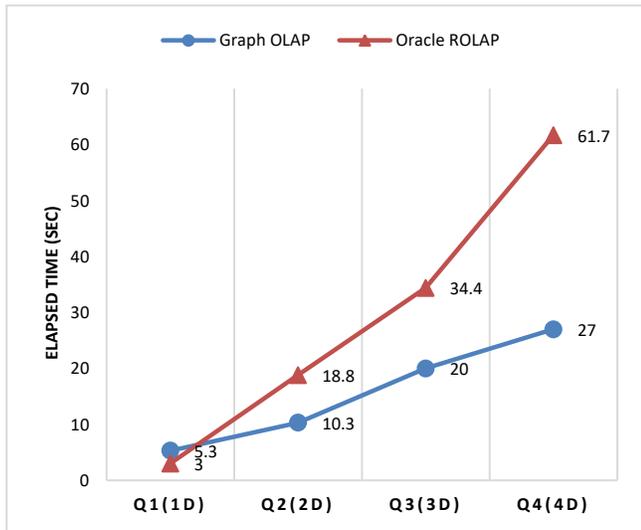


Fig. 7. Query Response Time by Dimensionality.

VI. CONCLUSION

The ability of graph technology to handle highly interconnected data makes it suitable for interactive analysis and more relevant for businesses today. In this paper, we addressed the topic of extending NoSQL graph-oriented databases to OLAP. We have proposed a modeling approach for implementing graph-based data warehouses using labeled nodes and edges. We have also shown how materialized aggregates can pre-computed across different levels to speed up query processing. At the physical level Neo4J engine is used as a graph-oriented database management system. Typical OLAP queries are rewritten using its declarative query language Cypher.

The Graph-OLAP implementation is compared to ROLAP one in terms of query performance and storage space, results show clearly that graph implementation of OLAP presents better performances than relational alternative in term of query response time when facing a huge data volume.

In the forthcoming extended work, we look forward to extending Cypher to support OLAP features by writing a user-defined aggregation function using the low-level API provided by Neo4J engine.

Without any doubt, using NoSQL technology to support OLAP features is a promising research direction. Therefore,

we claim that implementing OLAP engines under column-oriented and document-oriented databases using novel frameworks would be an interesting research issue that can be addressed.

REFERENCES

- [1] S. Chaudhuri and U. Dayal, "An overview of data warehousing and OLAP technology," *ACM SIGMOD Rec.*, vol. 26, no. 1, pp. 65–74, Mar. 1997, doi: 10.1145/248603.248616.
- [2] A. Nanda, S. Gupta, and M. Vijrania, "A Comprehensive Survey of OLAP: Recent Trends," in *2019 3rd International conference on Electronics, Communication and Aerospace Technology (ICECA)*, Coimbatore, India, Jun. 2019, pp. 425–430. doi: 10.1109/ICECA.2019.8822203.
- [3] A. Cuzzocrea, L. Bellatreche, and I.-Y. Song, "Data Warehousing and OLAP over Big Data: Current Challenges and Future Research Directions," in *Proceedings of the Sixteenth International Workshop on Data Warehousing and OLAP*, New York, NY, USA, 2013, pp. 67–70. doi: 10.1145/2513190.2517828.
- [4] M. Tremblay and A. Hevner, "Missing Data in OLAP Cubes: Challenges and Strategies," *J. Database Manag.*, vol. 32, p. 1, Jun. 2021, doi: 10.4018/JDM.2021070101.
- [5] K. Dehdouh, O. Boussaid, and F. Bentayeb, "Big Data Warehouse: Building Columnar NoSQL OLAP Cubes," *Int. J. Decis. Support Syst. Technol.*, vol. 12, no. 1, pp. 1–24, Jan. 2020, doi: 10.4018/IJDSST.2020010101.
- [6] K. Dehdouh, F. Bentayeb, O. Boussaid, and N. Kabachi, "Using the column oriented NoSQL model for implementing big data warehouses," *Int. Conf. Parallel Distrib. Process. Tech. Appl. PDPTA15*, pp. 469–475, 2015.
- [7] M. Boussahoua, O. Boussaid, and F. Bentayeb, "Logical Schema for Data Warehouse on Column-Oriented NoSQL Databases," in *Database and Expert Systems Applications*, vol. 10439, D. Benslimane, E. Damiani, W. I. Grosky, A. Hameurlain, A. Sheth, and R. R. Wagner, Eds. Cham: Springer International Publishing, 2017, pp. 247–256. doi: 10.1007/978-3-319-64471-4_20.
- [8] M. Chavalier, M. El Malki, A. Kopliku, O. Teste, and R. Tournier, "Document-oriented data warehouses: Models and extended cuboids, extended cuboids in oriented document," *Proc. - Int. Conf. Res. Chall. Inf. Sci.*, vol. 2016-Augus, 2016, doi: 10.1109/RCIS.2016.7549351.
- [9] Z. Challal, W. Bala, H. Mokeddem, K. Boukhalfa, O. Boussaid, and E. Benkhelifa, "Document-oriented versus Column-oriented Data Storage for Social Graph Data Warehouse," 2019, pp. 242–247. doi: 10.1109/SNAMS.2019.8931718.
- [10] C. Kamphuis, "Graph Databases for Information Retrieval," in *Advances in Information Retrieval*, Cham, 2020, pp. 608–612.
- [11] A. Bhattacharyya and D. Chakravarty, "(Graph Database: A Survey)," in *2020 International Conference on Computer, Electrical & Communication Engineering (ICCECE)*, Kolkata, India, Jan. 2020, pp. 1–8. doi: 10.1109/ICCECE48148.2020.9223105.
- [12] R. Kimball, "Kimball Dimensional Modeling Techniques," pp. 1–24, 2013, doi: 10.1016/B978-0-12-411461-6.00009-5.
- [13] F. Davardoost, A. Babazadeh Sangar, and K. Majidzadeh, "Extracting OLAP Cubes from Document-Oriented NoSQL Database Based on Parallel Similarity Algorithms," *Can. J. Electr. Comput. Eng.*, vol. 43, no. 2, pp. 111–118, 2020, doi: 10.1109/CJECE.2019.2953049.
- [14] S. Bouaziz, A. Nabli, and F. Gargouri, "Design a Data Warehouse Schema from Document-Oriented database," *Procedia Comput. Sci.*, vol. 159, pp. 221–230, 2019, doi: 10.1016/j.procs.2019.09.177.
- [15] E. Gallinucci, M. Golfarelli, and S. Rizzi, "Approximate OLAP of document-oriented databases: A variety-aware approach," *Inf. Syst.*, vol. 85, pp. 114–130, Nov. 2019, doi: 10.1016/j.is.2019.02.004.
- [16] M. L. Chouder, S. Rizzi, and R. Chalal, "EXODuS: Exploratory OLAP over Document Stores," *Inf. Syst.*, vol. 79, pp. 44–57, Jan. 2019, doi: 10.1016/j.is.2017.11.004.
- [17] A. Khalil and M. Belaissaoui, "New approach for implementing big datamart using NoSQL key-value stores," presented at the *Proceedings of 2020 5th International Conference on Cloud Computing and Artificial*

- Intelligence: Technologies and Applications, CloudTech 2020, Nov. 2020. doi: 10.1109/CloudTech49835.2020.9365897.
- [17] A. Khalil and M. Belaissaoui, "Key-value data warehouse: Models and OLAP analysis," presented at the 2020 IEEE 2nd International Conference on Electronics, Control, Optimization and Computer Science, ICECOCS 2020, Dec. 2020. doi: 10.1109/ICECOCS50124.2020.9314447.
- [18] H. Zhao and X. Ye, "A Practice of TPC-DS Multidimensional Implementation on NoSQL Database Systems," in Performance Characterization and Benchmarking, vol. 8391, R. Nambiar and M. Poess, Eds. Cham: Springer International Publishing, 2014, pp. 93–108. doi: 10.1007/978-3-319-04936-6_7.
- [19] C. Chen, X. Yan, F. Zhu, J. Han, and P. S. Yu, "Graph OLAP: a multi-dimensional framework for graph data analysis," Knowl. Inf. Syst., vol. 21, no. 1, pp. 41–63, Oct. 2009, doi: 10.1007/s10115-009-0228-9.
- [20] C. Chen, X. Yan, F. Zhu, J. Han, and P. S. Yu, "Graph OLAP: Towards Online Analytical Processing on Graphs," in 2008 Eighth IEEE International Conference on Data Mining, Pisa, Italy, Dec. 2008, pp. 103–112. doi: 10.1109/ICDM.2008.30.
- [21] P. Zhao, X. Li, D. Xin, and J. Han, "Graph cube: on warehousing and OLAP multidimensional networks," in Proceedings of the 2011 international conference on Management of data - SIGMOD '11, Athens, Greece, 2011, p. 853. doi: 10.1145/1989323.1989413.
- [22] C.-H. Chou, M. Hayakawa, A. Kitazawa, and P. Sheu, "GOLAP: Graph-Based Online Analytical Processing," Int. J. Semantic Comput., vol. 12, no. 04, pp. 595–608, Dec. 2018, doi: 10.1142/S1793351X18500071.
- [23] P. Wang, B. Wu, and B. Wang, "TSMH Graph Cube: A novel framework for large scale multi-dimensional network analysis," in 2015 IEEE International Conference on Data Science and Advanced Analytics (DSAA), Campus des Cordeliers, Paris, France, Oct. 2015, pp. 1–10. doi: 10.1109/DSAA.2015.7344826.
- [24] A. Castellort and A. Laurent, "NoSQL graph-based OLAP analysis," KDIR 2014 - Proc. Int. Conf. Knowl. Discov. Inf. Retr., pp. 217–224, 2014, doi: 10.5220/0005072902170224.
- [25] L. Gómez, B. Kuijpers, and A. Vaisman, "Performing OLAP over Graph Data: Query Language, Implementation, and a Case Study," in Proceedings of the International Workshop on Real-Time Business Intelligence and Analytics, Munich Germany, Aug. 2017, pp. 1–8. doi: 10.1145/3129292.3129293.
- [26] L. Gómez, B. Kuijpers, and A. Vaisman, "Online analytical processing on graph data," Intell. Data Anal., vol. 24, no. 3, pp. 515–541, May 2020, doi: 10.3233/IDA-194576.
- [27] A. Sellami, A. Nabli, and F. Gargouri, "Transformation of Data Warehouse Schema to NoSQL Graph Data Base," in Intelligent Systems Design and Applications, vol. 941, A. Abraham, A. K. Cherukuri, P. Melin, and N. Gandhi, Eds. Cham: Springer International Publishing, 2020, pp. 410–420. doi: 10.1007/978-3-030-16660-1_41.
- [28] M. Chevalier, M. El Malki, A. Kopliku, O. Teste, and R. Tournier, "Benchmark for OLAP on NoSQL technologies comparing NoSQL multidimensional data warehousing solutions," Proc. - Int. Conf. Res. Chall. Inf. Sci., vol. 2015-June, no. June, pp. 480–485, 2015, doi: 10.1109/RCIS.2015.7128909.
- [29] M. El Malki, A. Kopliku, E. Sabir, and O. Teste, "Benchmarking Big Data OLAP NoSQL Databases," in Ubiquitous Networking, vol. 11277, N. Boudriga, M.-S. Alouini, S. Rekhis, E. Sabir, and S. Pollin, Eds. Cham: Springer International Publishing, 2018, pp. 82–94. doi: 10.1007/978-3-030-02849-7_8.

Listing 8. Script loading in Neo4J

```
1. UNWIND ["sales-sf1.csv"] AS sourceFile
2. LOAD CSV WITH HEADERS
3. FROM "file:///" + sourceFile
4. AS row
5. FIELDTERMINATOR ';'
6. MERGE (cus:Customer {cname: row.c_name})
7. MERGE (r:Region {region: row.s_region_name})
8. MERGE (n:Nation {nation: row.s_nation_name})
9. MERGE (st:Store {store: row.s_name})
10. MERGE (n)-[:IN_REGION]->(r)
11. MERGE (st)-[:IN_NATION]->(n)
12. WITH date(row.o_orderDate) AS date,row,st,cus
13. MERGE (y:Year {year: toInteger(date.year)})
14. MERGE (q:Quarter {year: date.year, quarter: date.quarter})
15. MERGE (m:Month {year: date.year, month: date.month, quarter:date.quarter})
16. MERGE (q)-[:IN_YEAR]->(y)
17. MERGE (m)-[:IN_QUARTER]->(q)
18. MERGE (b:Brand {brand: row.p_brand})
19. MERGE (prod:Product {product: row.p_name})
20. MERGE (prod)-[:OF_BRAND]->(b)
21. WITH
22. st, m, prod, row,cus,
23. st.store + '_' + toString(m.year) + '_' + toString(m.month) + '_' + prod.product+ '_' +
   cus.cname AS SalesID
24. MERGE (f:Sales {fid: SalesID})
25. ON CREATE
26. SET f.sales = toFloat(row.sales),
27. f.quantity = toInteger(row.quantity)
28. ON MATCH
29. SET f.sales = f.sales + toFloat(row.sales),
30. f.quantity = f.quantity + toInteger(row.quantity)
31. MERGE (f)-[:IN_STORE]->(st)
32. MERGE (f)-[:IN_MONTH]->(m)
33. MERGE (f)-[:FOR_PRODUCT]->(prod)
34. MERGE (f)-[:BY_CUSTOMER]->(cus);
```

Sentiment Analysis to Explore User Perception of Teleworking in Saudi Arabia

Malak Nazal Alotaibi, Zahyah H. Alharbi

Management Information Systems Department

College of Business Administration, King Saud University, Riyadh, Saudi Arabia

Abstract—Due to the emergence of the COVID-19 pandemic in 2019, many public and private organizations from different sectors in Saudi Arabia were forced to enforce teleworking as the main work arrangement. This paper seeks to understand the experience and attitude of the public toward remote work by analyzing Twitter data from March 2020 to July 2021 by using "Mazajak" the online Arabic analyzer. A corpus of 39,523 tweets with hashtags mentioning the teleworking program in Saudi Arabia was obtained. The results indicate that neutrality was the most prevalent sentiment with 58.21%, followed by positive sentiment with 30.67%. Thematic analysis was used to identify themes in the tweets with positive and negative sentiment. Flexibility, teamwork, teleworking preference, and learning were the major themes related to positive sentiment, while themes related to negative sentiment were private sector, companies, and fake.

Keywords—Mazajak; sentiment analysis; thematic analysis; telework; remote work

I. INTRODUCTION

Recently, the concept of teleworking or work from home (WFH) has been gaining in popularity due to the advancement in information technology and the emerging tools and applications for communication and task management. There are many definitions of teleworking, one of which views teleworking as a type of "work arrangement with high flexibility in which employees perform all or a substantial part of their work physically separated from the location of their employer, using IT for communication and operations" [1].

In 2020, a global public health crisis emerged due to the spread of the COVID-19 virus. The ensuing pandemic forced many national governments worldwide to adopt lockdown policies. Social distancing became mandatory, which compelled organizations to transition their employees away from regular work and toward telework, the aim being to protect employees and reduce the societal impact of the virus. These radical changes in the work environment have had both advantages and disadvantages. Ipsen et al. investigated the advantages and disadvantages of teleworking during the pandemic in Europe, classifying them into the following six categories: work-life balance, work efficiency, and work control (as the advantages), and home office constraints, work uncertainties, and inadequate tools (as the disadvantages) [2].

Before the COVID-19 pandemic in Saudi Arabia, firms from different sectors had limited experience with teleworking as a work arrangement. However, the situation after the pandemic in Saudi Arabia was – and continues to be – similar

to other countries around the world. In March 2020, teleworking was enforced on firms and organizations from different sectors as part of the precautionary measures to combat the COVID-19 pandemic. As a result, the pandemic gave rise to a forced experiment that led to significant changes in work methods across Saudi Arabia.

Several years before the pandemic, in 2016, the Saudi Ministry of Human Resources and Social Development launched a teleworking initiative [3], the aim of which was to fill the geographical gap between employers and job seekers based on the use of electronic work environment [4]. Despite the initiation of this teleworking program, teleworking was not popular prior to the pandemic. Understandably, its popularity has increased dramatically in Saudi Arabia in the wake of the ongoing pandemic.

In this research, we conducted a preliminary study in which 187 employees were targeted from different sectors. The sample consisted of 79.8% females and 20.2% males, 68.6% of whom were aged between 20 and 40 years. 30.3% of the sample was aged 41 to 60 years, and only 1% were older than 60. The majority of the participants (68.6%) held a bachelor's degree and 19.1% have a master's degree. The sectors represented in the sample were the educational sector (37.2%), public sector (32.4%), private sector (26.1%), and healthcare sector (4.3%). The results of this study show that 63.8% of the participants had never tried teleworking before the pandemic, which indicates that the teleworking concept is regarded as a relatively novel concept in the Saudi labor market.

This study's preliminary results suggest that the pandemic has prompted Saudi organizations to implement teleworking as a way of adapting to the global changes. These changes in the work dynamics around the world can be characterized as a double-edged sword in that they may affect both organizations and employees positively or negatively. This can cause different perspectives on the impact of teleworking.

Studying Public perspectives on the teleworking experience during the pandemic is an important factor for analyzing the overall experience and learning how to improve it. Furthermore, knowing the reasons behind the negative or the positive perspective of workforces is expected to aid businesses in utilizing teleworking to maximize its benefits. One way to analyze people's opinions toward a topic is through sentiment analysis. Sentiment is defined as the emotion behind a mention of a certain topic, brand, or service in the social universe. It is a way of gaining an understanding

of the feeling surrounding a topic, brand, or service [4]. A popular way to analyze sentiment toward a subject is by analyzing their social media posts, such as tweets, Facebook posts, and Instagram posts. Although there are several social media platforms for posting and sharing comments, Twitter was targeted in this paper. Twitter is an informal microblogging social media platform that allows users to share posts of up to 280 characters called “tweets” [5]. It provides an accessible and large amount of data, particularly given that there was an average of 206 million daily active users in 2021 [6]. Consequently, the large, timely, rich, and easily accessible data available on Twitter were the reasons for choosing this social media platform as the data source for this paper.

Despite the importance of analyzing the teleworking experience, the situation in Saudi Arabia has not been studied, to the best of the researcher's knowledge. This study aims to fill this gap by answering the following question: What is society's perspective and the public's opinion toward the teleworking experience in Saudi Arabia? In addition, this study seeks to achieve the following objectives:

- To investigate the teleworking experience in Saudi Arabia from the societal perspective.
- To explore the themes that correlate with positive and negative opinions using sentiment analysis, visualization, and thematic analysis.

The remainder of this paper is organized as follows. Section 2 presents related work in the form of a literature review. Section 3 introduces the methodology used in this paper to collect and analyze data. Section 4 discusses the results of the sentiment and thematic analysis. Section 5 discusses the results and presents recommendations. Lastly, Section 6 draws concluding remarks for this study.

II. LITERATURE REVIEW

Previous studies on the concept of telework or WFH were mostly performed before the COVID-19 outbreak. More recent studies on the concept cover the impact of COVID-19 on teleworking using different techniques, including sentiment analysis. In this section, a review of the literature is given focusing on the concepts of telework, sentiment analysis, and Arabic sentiment analysis.

A. Telework

In 1973, Nilles [7] coined the term “telework”. The author identified telework as a work arrangement that “includes all work-related substitutions of telecommunications and related information technologies for travel”. The evolution of teleworking has since progressed through three generations [8]. The first generation is the home office, where employees exclusively work from home. The second generation is the mobile office, where working hours were partially replaced with teleworking. The third generation is the virtual office, which refers to an informal work arrangement that has fewer regulations than the usual work arrangement. Currently, with the use of smartphones and advanced technologies, organizations are part of the third generation of teleworking. Employees who work from home (i.e., teleworkers) are classified into three main types, according to Vries et al [1].

The first type is home-based teleworkers, where employees perform their tasks at home. The second type is teleworking from remote offices, where employees perform their tasks in an office that is remote from the main office. The third type is mobile telework, which includes employees who need to travel or visit customers as part of their job. Briefly, teleworking is a spectrum of practices, not a homogeneous entity [9].

The negative and positive effects of telework have been the subject of many research projects. Teleworking can impact employees, management, and the organization as a whole. Telework can affect employees' effectiveness as well as their personal life. For example, Adamovic [10] conducted a study to investigate the effect of teleworking on job stress and telework effectiveness. The study's hypotheses were tested using a three-phase survey targeting 604 teleworkers from different countries and a variety of organizations in different sectors. The survey measured cultural values, beliefs toward telework, and job stress. The results showed that telework only reduces job stress when employees do not believe that teleworking leads to social isolation. Furthermore, the results indicated that employees with high power distance scores typically held negative beliefs about telework, whereas employees with high individualism scores tended to have strongly positive beliefs regarding the effectiveness of telework.

An equally significant aspect of teleworking is its impact on employees' quality of life, which was studied by Nedelcu in [11]. Nedelcu conducted a survey with 261 employed undergraduate and graduate students from Nicolae Titulescu University in Bucharest. The results showed that telework can positively impact the quality of employees' personal and professional life due to the following factors: reduction of work-related stress, increasing employee autonomy, reducing costs related to work, and increasing motivation and commitment. Conversely, the study indicated that 73% of dissatisfied employees agreed that their dissatisfaction with telework experience was because they were unable to separate their personal time from their professional time. Moreover, 86% of the respondents agreed that their dissatisfaction was related to the loss of social interactions and team spirit. Similarly, Golden in [12] stated that telework can affect employees' commitment, exhaustion, and turnover intention. The researcher used a sample of 393 teleworkers in a large firm in the USA. The results indicated that teleworking increases employees' commitment and lowers turnover intention.

While many previous studies have targeted employees from all levels, Silva et al. [13] focused on managers and their attitudes toward telework. The study used the technology acceptance model (TAM) as a theoretical framework and the collected data were analyzed using structural equation modeling (SEM). The results showed that the attitude of managers toward telework adoption was influenced by improvements in information security tools, employees' self-efficacy beliefs, and managerial practices.

Although many studies have investigated the effect of telework on different aspects of organizations and employees,

several recent studies have examined teleworking during the COVID-19 pandemic. For example, Belzunegui-Eraso et al. [14] and Ipsen et al. [2] both found that telework working conditions during the COVID-19 pandemic have mostly been positive. According to these researchers, the principal advantages of teleworking during the COVID-19 pandemic are work-life balance, work efficiency, and work control.

B. Sentiment Analysis

To the best of the researcher's knowledge, sentiment analysis has only been applied in two studies focusing on telework, both of which were undertaken during the time of the COVID-19 pandemic. The first study, Zhang et al. [15], aimed to understand the attitudes of the general public toward teleworking by analyzing tweets using natural language processing (NLP) techniques. The results uncovered themes among tweets including mental health, teamwork, leadership, and work-life balance. The other study was conducted by Goyal and Malhotra [16], and while it was similar to [15] regarding its objective and the use of tweets as the data source, the main difference is that [16] used the R programming language as the method for sentiment analysis. Both studies showed that public attitudes toward working from home have generally been positive.

C. Arabic Sentiment Analysis

Sentiment analysis for the Arabic language is different from English. One of the main reasons for this is that the Arabic language is more complex than English, richer in terms of its morphology, and has different dialects. In each dialect, the meaning of the words can be very different, which makes sentiment analysis for the Arabic language more challenging. The literature on Arabic sentiment analysis offers many ways to handle the challenging aspects of analyzing the Arabic language.

In [17], the researchers introduced an aspect-based sentiment analysis system for Arabic reviews of hotels. They implemented a deep recurrent neural network (RNN) along with a support vector machine (SVM) approach, and the results showed that the SVM approach outperformed the deep RNN approach. Moreover, in [18], Alayba, Palade, et al built a system for sentiment analysis based on the integration of convolutional neural networks (CNNs) and long short-term memory (LSTM) networks. The hybrid model was tested on two Arabic datasets and achieved high classification accuracy.

In a more recent study, Al-Alfy & Al-Azany in [19] had two goals: first, to compare the performance of machine learning algorithms in discovering the polarity of Arabic tweets using neural word embedding as the feature extractor; and second, to examine the effect of different oversampling techniques in handling the imbalanced nature of the data. The paper's findings were that the geometric mean (GM) attained its highest value when the stochastic gradient descent (SGD) classifier was used with oversampling. The GM improved in all cases when combined with oversampling except for the Gaussian naïve Bayes (GNB) and random forest (RF) classifiers.

A notable tool for Arabic sentiment analysis is Mazajak, which is the first online Arabic sentiment analyzer. In [20],

Abu Farah & M proposed this online system, which analyzes Arabic sentiment using a deep learning model. Mazajak achieved state-of-the-art results on many Arabic dialects. The system's results are based on the benchmark of three Arabic datasets: the SemEval 2017 task, ArSAS, and ASTD. Given that Mazajak has outperformed all other Arabic sentiment analysis systems, this paper utilizes the Mazajak system as a tool for sentiment analysis, as described in the next section.

III. METHODOLOGY

This section present the research methodology, which consisted of four phases: data collection and preparation, data analysis and performance evaluation.

A. Data Collection and Preparation

In this study, a dataset consisting of 39,523 tweets was retrieved from three hashtags (#_تعلیق_العمل, #_تعلیق_العمل_في_السعودية, #_العمل_عن_بعد). In addition to the hashtags, the tweets from the telework program account mention were retrieved for the period from January 2019 to July 2021. The details of the retrieved data are shown in Table I.

TABLE I. TWEET CLASSIFICATION

Sources	Number of Retrieved Tweets
@TeleworksKSA mention	8,235 tweets
#Telework	17,018 tweets
#Work_suspension_SaudiArabia	3,611 tweets
#Work-Susbension	10,659 tweets

Data preprocessing was conducted using RapidMiner, a Java-based open-source software that provides several operators for text processing. These operators include tokenization, stemming, and stopword filtering [21]. Furthermore, RapidMiner supports multiple languages, one of which is Arabic.

The first step was to integrate the four datasets shown in Table I into one large dataset. Following this, a data reduction phase was performed to reduce or eliminate noisy data (e.g., meaningless or irrelevant tweets). its involved scanning the data for specific words and deleting any tweets containing these words. For example:

Offers- designs- IELTS-CV-Marriage

الفاتورة – مسيار – ايلتس- سيرة ذاتية – تصميم – منتجات – عروض

After this step, the dataset was reduced to 22,587 tweets. To prepare the data, several data cleaning techniques were used such as normalization, transformation, tokenization, and stopword filtering. First, RapidMiner was used to clean the data in terms of links, http tags, and other symbols such as @, #, and [] using the Replac operator. The next process was data normalization, the goal of which is to reduce the complexity of a text and transform it into a simpler form. For example, the letters (ا , و , ي) were converted to (آ , و , ي). In turn, data transformation was applied. According to [22], in this preprocessing step, the data is converted so that the mining process result can be applied in a more efficient way. In the process of filtering stopwords, unimportant words such

as (أيضا, إلى, ثم) are removed from the data. RapidMiner supports this function for the Arabic language. The second process was tokenization, wherein each sentence is split into several words called tokens.

After tokenization, an operator for filtering tokens by length was used to filter any words with fewer than 4 characters or more than 25 characters. Stemming is considered an important step in the transformation process for English data to convert the words into their roots. This step was not used in this research since the dataset was in Arabic and stemming can negatively affect the classifier's overall performance and accuracy [23]. Instead of stemming, another technique was used to improve sentiment prediction: namely, replacing synonyms with a single word that have the same meaning. This improved the prediction because it made the dataset more unified and easier to process. An example is replacing the phrase "what is the solution", which is "وَشِ الحَلُّ / ما هو الحل" in Arabic with the word "problem", which is "مشكلة". Another example is replacing the Arabic phrase for "useless", which is "لا فائدة" with the word "عبث", which is the one-word Arabic synonym of the phrase.

B. Data Analysis

Sentiment analysis, also called opinion mining, is "a field of study that analyzes people's opinions, attitudes, emotions, and sentiments toward topics, events, products, or services, and their attributes" [24]. Sentiment analysis is a term used interchangeably with opinion extraction, sentiment mining, emotion analysis, and affect analysis. Sentiment analysis includes several steps and the final step is finding the polarity of the data [25]. A popular way to analyze people's sentiment toward a subject is by analyzing their social media posts. There are three main methods for analyzing sentiment in social media content: the machine learning approach, the lexicon approach, and the hybrid approach. The machine learning approach is divided into supervised learning and unsupervised learning, while the lexicon approach is divided into the dictionary-based approach and corpus-based approach. In supervised learning, classification models are used such as the naïve Bayes, Bayesian network, SVM, artificial neural network (ANN), and decision tree [26].

This study leveraged the existence of Mazajak, an online Arabic sentiment analyzer, due to its remarkable performance and state-of-the-art results [27]. Mazajak uses NLP and machine learning to classify tweets into three categories: positive, negative, or neutral. The model used in the system was built on a Convolutional Neural Network (CNN) that works as the feature extractor, after which the embeddings are fed into the max-pooling layer. In turn, the extracted features are fed into an Long Short Term Memory network (LSTM), taking into consideration the context and order of the words. The LSTM is followed by a softmax layer, which produces the output classes [20].

C. Performance Evaluation

To evaluate the performance of the online tool, we used three performance parameters: accuracy, precision, and recall.

Accuracy is calculated as the correct classifications divided by all classifications. In this study, it was calculated using Equation (1).

$$Accuracy = \frac{TP+TN+TNe}{TP+TN+TNe+FP+FN+FNe} \quad (1)$$

In the numerator of Equation (1), TP is the number of true positive predictions, TN is the number of true negative predictions, and TNe is the number of true neutral predictions. In addition, in the denominator, FP is the number of false positive predictions, FN is the number of false negative predictions, and FNe is the number of false neutral predictions.

The second measure is precision, which is also known as the positive predictive value. It measures the classifier's performance in predicting the true polarity of the test dataset [28]. Precision was calculated in this study as shown in Equation (2).

$$Precision = \frac{TP}{TP+FP} \quad (2)$$

The third measure is recall, which measures the sensitivity of the classifier. This was calculated using Equation (3).

$$Recall = \frac{TP}{TP+FN} \quad (3)$$

The results of the previous measures were high. In particular, the value of the accuracy was 88.04%, which indicates a high accuracy rate for the tool. Moreover, the precision and recall values were high at 0.97 and 0.88, respectively. Hence, the tool can be used to analyze the data.

IV. FINDINGS

A. Sentiment Analysis

The results of the sentiment analysis indicate that 58.4% of the tweets were classified as neutral. The percentage of positive tweets exceeded the negative tweets by more than the double, with 30.76% positive and 11.15% negative, as shown in Fig. 1.

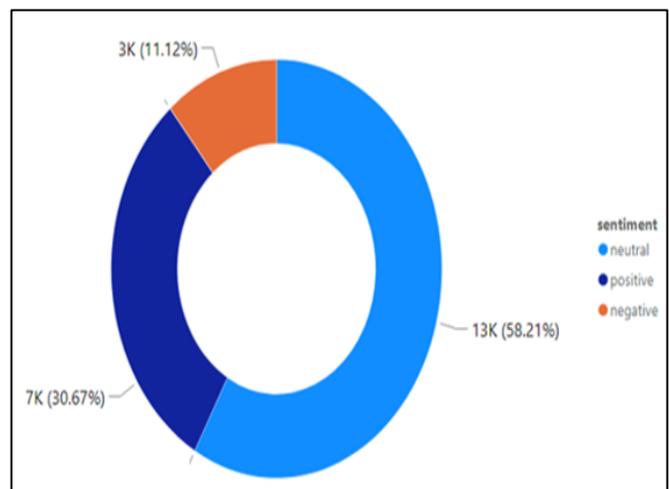


Fig. 1. Data Polarity.

The frequencies of the words in the tweets were analyzed using a word cloud. Fig. 2 shows the results of the word cloud for the most frequently used words in the dataset.

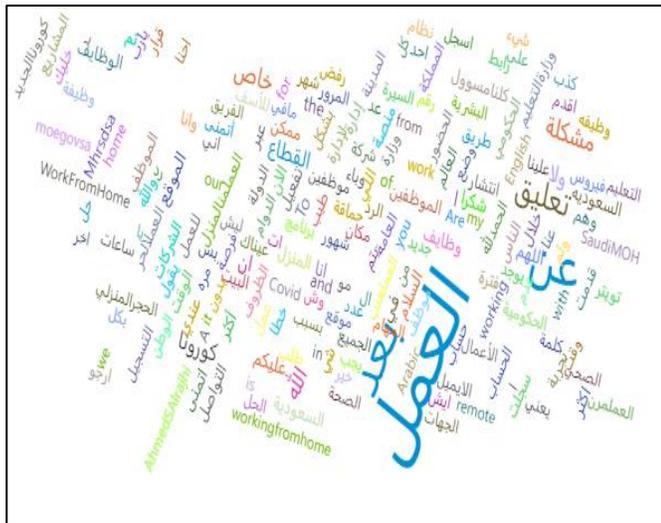


Fig. 2. Word Cloud for the Dataset.

B. Thematic Analysis

To identify the factors that correlated with the polarity of the tweets, separate word clouds were created for the positive and negative tweets. However, it was necessary to exclude some words such as (كورونا, سلمان, وباء) which translates to (Corona, Salman, Epidemic). The rationale for excluding these words is that they negatively impacted the results. The results for both the word clouds are shown in Fig. 3 and Fig. 4.



Fig. 3. Positive Word Cloud.

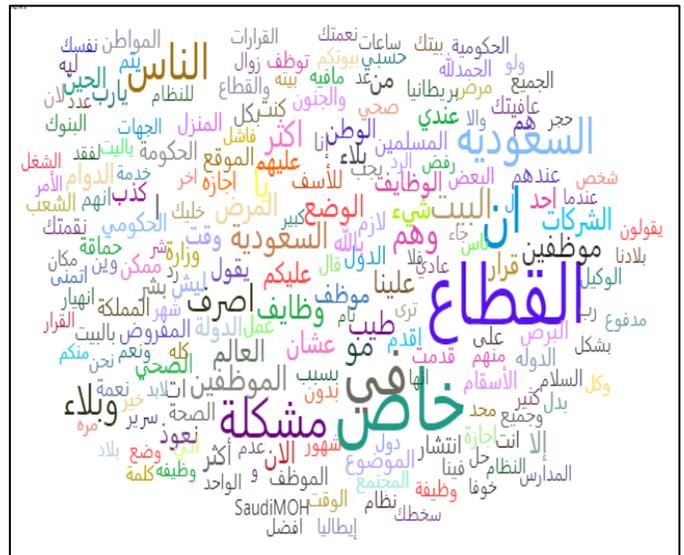


Fig. 4. Negative Word Cloud.

The positive word cloud contained words such as “thank you”, “beautiful”, “team”, “home”, “better”, “technology”, “experience”, “communication”, “flexibility”, and “productivity”. On the other hand, the negative word cloud contained words such as “private”, “sector”, “problem”, “companies”, “sickness”, and “delusion”. From the previous figures, we can extract some themes that correlated with positive or negative opinions. The word “productivity” was one of the most frequently used words in positive tweets. Furthermore, its use was often accompanied by words such as “effectiveness” and “efficiency”. Examples of these tweets are:

“We left our offices but we increased our productivity and efficiency through trusting our employees.”

“Teleworking is a wonderful arrangement to avoid traffic and improve productivity, regardless of its challenges.”

“Teleworking can increase the team’s productivity if it is managed effectively and efficiently.”

“Teleworking taught us that productivity is about the quality rather than the quantity of work, and we have to consider reducing the number of working hours.”

Another theme was the comparison between the teleworking work arrangement and the regular work arrangement. One of the most frequently used words was “better”, which was often used to describe health, work, accomplishments, and standards during remote work. Examples of this theme are as follows:

“I noticed that transforming to teleworking maximizes the work output and allows everybody to communicate better in a shorter time; it is a better experience.”

“Having more breaks when teleworking contributes to better health.”

"As a teleworker, I think that work standards and accomplishments are better when teleworking – it's the right option for a better environment, and in the future, teleworking will be part of the digital transformation."

"Teleworking reduces costs and gives organizations an opportunity to hire better employees regardless of where they live."

The word "learning" was another common word used in the positive dataset. Teleworking was a new experience for many employees during the COVID 19 pandemic, which forced them to learn about new technologies and applications. Thus, "technology" was also a common word. Examples include the following:

"Teleworking has many advantages such as flexibility and productivity, and it is a great opportunity to learn new habits and skills."

"We are forced to go through this experience and we need patience and creativity to use it to learn."

"The teleworking experience is an enriching experience that has given us the opportunity to learn new technologies."

The word "team" was a popular word in the positive dataset. Tweets mentioning this word talked positively about the creativity of the team arising from telework, as well as trust and communication. The following are examples of this theme:

"A successful team during the period of telework is marked by trust, respect, and helping each other."

"Teleworking experience is fun and special, especially when you have the technology and tools to help and, most importantly, a creative and passionate team."

"A lesson we can learn from the teleworking experience is that building trust in your team is the greatest incentive for productivity."

"Your creative and professional team is the most important asset – communicate with them and encourage them to maintain their efficiency and productivity during working from home."

Lastly, working with flexible hours correlated with positive opinions about teleworking as it is considered one of the advantages of this work arrangement. Some of the tweets that mentioned flexibility are the following:

"Teleworking is flexible, which makes employees more productive and happier; it opens a new gateway to opportunities."

"The flexibility of teleworking gives us a better work-life balance."

"The idea of working from home is so much better and more flexible, which makes it easier."

"Having flexible working hours contributes to better health and life quality for employees."

On the other hand, two of the most common words in the negative dataset were "private" and "sector", which occurred

in 24.58% of the tweets in the negative dataset. Most of these tweets highlighted how the private sector continued to work as normal during the pandemic, whereas employees in the public sector were teleworking. In addition, the word "companies" (or "company") was mentioned frequently in negative tweets. Some examples are given as follows:

"The company that I work for misunderstands the concept of teleworking. Managers ask us to hold meetings and perform tasks all day – even at night – and on top of that, they've cut our salaries."

"Why are private-sector employees being forced to go into the office during the pandemic? We are humans too and we can get sick."

"Why doesn't the work suspension apply to private-sector employees? We can also get the virus and put our families in danger."

In addition, words such as "fake", "failed", and "problem" were common in the negative tweets, especially in tweets that mentioned the teleworking program. Specifically, the words "Fake", "Lie", and "Problem" were present in 14.09% of the negative dataset, and 12.52% explicitly mentioned the teleworking program.

In the tweets that mentioned the teleworking program, 39.41% were negative and only 13.67% were positive. The word "problem" was featured in 32.97% of the tweets that identified a problem with the website and described it as a failed system. The words "fake" and "lie" were present in 16.21% of the tweets, where people accused the program of being fake or offering fake job opportunities. Notably, this was the most common theme in tweets mentioning the name of the program. The following are some of the instances of this theme:

"Are you sure that the website is working? I applied for tons of jobs and I think it is all fake because the organizations ignored us and didn't even send an apology."

"This website is a lie – it's offering fake job opportunities."

"I applied for jobs aligned with my CV and nobody contacted me – I'm starting to think that it's all fake."

"The website is very slow and has many problems."

V. DISCUSSION AND RECOMMENDATIONS

From the results discussed previously, the first important topic we investigated was the overall experience of teleworking. The public orientation toward neutrality and positivity may be due to the national lockdown, which helped employees to focus more on job tasks and be less stressed about balancing work and social life. This led to many not viewing the situation as a negative experience.

Since the pandemic, the use of teleworking has increased and, as a result, work flexibility has become more common. This has elevated the importance for organizations to build a positive teleworking experience to achieve high productivity levels. Offering flexible hours for teleworkers can improve their productivity. In a study conducted in Japan, the

researchers found that suitable telework hours increased employee productivity, whereas long working hours decreased productivity [29]. For the teleworking program, the results showed that it is not a feasible program and it has not been well received by the public. We recommend a more transparent recruitment process and a user-friendly website to avoid problems with registering and applying for jobs [30].

VI. CONCLUSION AND FUTURE WORK

This paper performed sentiment analysis and thematic analysis focusing on the teleworking experience and teleworking program in Saudi Arabia. The purpose of this study was to gain insight into the experience of teleworking among the public in Saudi Arabia during the work suspension period, as well as to determine the public sentiment of the experience. Furthermore, this study sought to analyze public sentiment on the teleworking program. A dataset extracted from Twitter, consisting of tweets posted between March 2020 and July 2021, was used to analyze the public sentiment toward teleworking. Mazajak, an online Arabic sentiment analyzer, was used for data analysis. The results revealed that neutrality was the most prevalent sentiment in the tweets, followed by positive sentiment. Furthermore, the results detected themes of flexibility, teamwork, teleworking preference, and learning in association with the positive view of the practice. By contrast, the themes related to negative sentiment were the private sector, companies, and fake.

Future work should concentrate on comparing the public view of teleworking before and after the pandemic. In addition, it would be beneficial to study the emerged factors that affect the success of teleworking in Saudi Arabia to improve the experience and leverage its advantages. Finally, this research used data from public users on twitter with different economical and educational backgrounds. However, focusing on the impact of different demographics on the effectiveness and acceptance of teleworking deserve further research.

REFERENCES

- [1] Vries, H. D., Tummers, L., & Bekkers, V. (2018). The Benefits of Teleworking in the Public Sector: Reality or Rhetoric? *Review of Public Personnel Administration*, 39(4), 570-593. doi:10.1177/0734371x18760124.
- [2] Ipsen, C., van Veldhoven, M., Kirchner, K., & Hansen, J. P. (2021). Six key advantages and disadvantages of working from home in Europe during COVID-19. *International Journal of Environmental Research and Public Health*, 18(4), 1826. https://doi.org/10.3390/ijerph18041826.
- [3] The launch of the "Teleworking" program to maximize job opportunities. *جريدة الرياض*. (2016, January 23). Riyadh newspaper authors. (2016). Launch of the "Telecom work" program to increase career opportunities. Retrieved December 22, 2021, from https://www.alriyadh.com/1121979.
- [4] Duwairi, R. M., Marji, R., Sha'ban, N., & Rushaidat, S. (2014). Sentiment analysis in Arabic tweets. 2014 5th International Conference on Information and Communication Systems (ICICS). https://doi.org/10.1109/iacs.2014.6841964.
- [5] Telework program. (n.d.). Retrieved November 2, 2021, from https://teleworks.sa/ar/about-us/.
- [6] Techopedia. (2013, January 10). What is Twitter? - definition from Techopedia. Techopedia.com. Retrieved December 22, 2021, from https://www.techopedia.com/definition/4957/twitter Published by Statista Research Department, & 19, N. (2021, November 19). Twitter: Most users by country. Statista. Retrieved December 22, 2021, from https://www.statista.com/statistics/242606/number-of-active-twitter-users-in-selected-countries/.
- [7] Jackson, P., & der, W. J. M. M. van. (1998). FROM TELEWORKING TO NETWORKING Definitions and trends. In *Teleworking: International perspectives: From telecommuting to the Virtual Organisation*. essay, Routledge.
- [8] Messenger, J. C., & Gschwind, L. (2016). Three generations of telework: New icts and the (r)evolution from Home Office to Virtual Office. *New Technology, Work and Employment*, 31(3), 195-208. https://doi.org/10.1111/ntwe.12073.
- [9] Collins, M. (2005). The (not so simple) case for teleworking: A study at Lloyd's of London. *New Technology, Work and Employment*, 20(2), 115-132. https://doi.org/10.1111/j.1468-005x.2005.00148.x.
- [10] Adamovic, M. (2022). How does employee cultural background influence the effects of telework on job stress? the roles of power distance, individualism, and beliefs about telework. *International Journal of Information Management*, 62, 102437. https://doi.org/10.1016/j.ijim.2021.102437.
- [11] Nedelcu, E. (2020). The Perspective of Young People on the Effects of Telework on the Quality of Life at Work. *Romanian Review of Social Sciences*, 10(19), 3-12.
- [12] Golden, T. D. (2006). Avoiding depletion in virtual work: Telework and the intervening impact of work exhaustion on commitment and turnover intentions. *Journal of Vocational Behavior*, 69(1), 176-187. doi:10.1016/j.jvb.2006.02.003.
- [13] Silva-C, A., R, I. A., & A, J. A. (2019). The attitude of managers toward telework, why is it so difficult to adopt it in organizations? *Technology in Society*, 59, 101133. doi:10.1016/j.techsoc.2019.04.009.
- [14] Belzunegui-Eraso, A., & Erro-Garcés, A. (2020). Teleworking in the context of the COVID-19 crisis. *Sustainability*, 12(9), 3662. https://doi.org/10.3390/su12093662.
- [15] Zhang, C., Yu, M. C., & Marin, S. (2021). Exploring public sentiment on enforced remote work during COVID-19. *Journal of Applied Psychology*, 106(6), 797-810. https://doi.org/10.1037/apl0000933.
- [16] Malhotra, N., & Goyal, T. (2021). Sentiment analysis using Twitter information flow about the new education policy introduced in India in 2020. *INTERNATIONAL JOURNAL OF MANAGEMENT*, 11(12). https://doi.org/10.34218/ijm.11.12.2020.228.
- [17] Al-Smadi, M., Qawasmeh, O., Al-Ayyoub, M., Jararweh, Y., & Gupta, B. (2018). Deep recurrent neural network vs. support vector machine for aspect-based sentiment analysis of Arabic hotels' reviews. *Journal of Computational Science*, 27, 386-393. https://doi.org/10.1016/j.jocs.2017.11.006.
- [18] Alayba, A. M., Palade, V., England, M., & Iqbal, R. (2018). A combined CNN and LSTM model for Arabic sentiment analysis. *Lecture Notes in Computer Science*, 179-191. https://doi.org/10.1007/978-3-319-99740-7_12.
- [19] El-Alfy, E.-S. M., & Al-Azani, S. (2020). Empirical study on imbalanced learning of Arabic sentiment polarity with neural word embedding. *Journal of Intelligent & Fuzzy Systems*, 38(5), 6211-6222. https://doi.org/10.3233/jifs-179703.
- [20] Abu Farha, I., & Magdy, W. (2019). Mazajak: An online Arabic sentiment analyzer. *Proceedings of the Fourth Arabic Natural Language Processing Workshop*. https://doi.org/10.18653/v1/w19-4621.
- [21] Duwairi, R. M., & Qarqaz, I. (2014). Arabic sentiment analysis using supervised classification. 2014 International Conference on Future Internet of Things and Cloud. https://doi.org/10.1109/ficloud.2014.100.
- [22] García, S., Luengo, J., & Herrera, F. (2014). Data reduction. *Intelligent Systems Reference Library*, 147-162. https://doi.org/10.1007/978-3-319-10247-4_6.
- [23] Wahbeh, A., Al-Kabi, M., Al-Radaideh, Q., Al-Shawakfa, E., & Alsmadi, I. (2011). The effect of stemming on Arabic text classification. *Information Retrieval Methods for Multidisciplinary Applications*, 207-225. https://doi.org/10.4018/978-1-4666-3898-3.ch013.
- [24] Liu, B. (2012). Sentiment analysis and opinion mining. Morgan & Claypool.
- [25] Sharma, D., Sabharwal, M., Goyal, V., & Vij, M. (2019). Sentiment analysis techniques for Social Media Data: A Review. *First International*

- Conference on Sustainable Technologies for Computational Intelligence, 75–90. https://doi.org/10.1007/978-981-15-0029-9_7.
- [26] Sharma, D., Sabharwal, M., Goyal, V., & Vij, M. (2019). Sentiment analysis techniques for Social Media Data: A Review. First International Conference on Sustainable Technologies for Computational Intelligence, 75–90. https://doi.org/10.1007/978-981-15-0029-9_7.
- [27] Albalawi, Y., Buckley, J., & Nikolov, N. S. (2021). Investigating the impact of pre-processing techniques and pre-trained word embeddings in detecting Arabic health information on social media. *Journal of Big Data*, 8(1). <https://doi.org/10.1186/s40537-021-00488-w>.
- [28] Raut, A., & Pandey, R. K. (2019). Sentiment Analysis using Optimized Feature Sets in Different Twitter Dataset Domains. *International Journal of Innovative Technology and Exploring Engineering*, 8(11), 3035–3039. <https://doi.org/10.35940/ijitee.k2195.0981119>.
- [29] Kazekami, S. (2020). Mechanisms to improve labor productivity by performing telework. *Telecommunications Policy*, 44(2), 101868. <https://doi.org/10.1016/j.telpol.2019.101868>.
- [30] Thompson, L. F., Braddy, P. W., & Wuensch, K. L. (2008). E-recruitment and the benefits of organizational web appeal. *Computers in Human Behavior*, 24(5), 2384-2398.

Framework for Development of 3D Temple Objects based on Photogrammetry Method

Herman Tolle*, Ratih Kartika Dewi, Komang Candra Brata, Benyamin Perdamean
Research Group of Media, Game and Mobile Technology, Computer Science Faculty
Brawijaya University, Malang, Indonesia

Abstract—Indonesia has a lot of cultural buildings that need to capture as digital objects for other purposes, especially for digital preservation. One of the methods of making 3D objects is using the photogrammetry method. The photogrammetry method makes 3D objects using many photos captured by the camera that will be integrated into the software and processed into 3D objects. In this research, the framework for modeling 3D objects of the cultural building is proposed based on the photogrammetry approach. This framework includes data capture, modeling and processing, and calibration. This framework was tested while making the 3D object of the Candi Badut temple and reported that the 3D model has significantly had similarities with the real object. The framework is useful for standard guidelines for making 3D modeling of historical relics efficiently.

Keywords—Photogrammetry; historical relic; 3D objects; digital preservation; virtual reality

I. INTRODUCTION

Indonesia has many cultures and ancient historical relics that need to be preserved. Digitization is one of the efforts that can be done to preserve historical buildings. One of the historical relics found in the Malang Regency is the Badut Temple. The Badut Temple is located in Karangbesuki, Dau District, Malang Regency, East Java Province, about 10 kilometers from Malang City. In Indonesia, the role of the temple is very important in the introduction of the cultural history of Indonesia [1]. Most of the temple has reliefs on its wall containing a series of important stories to capture and create lessons for the current generation.

By digitizing 3D objects, we can preserve the visual form of a historical object in more detail. The Building Information Modeling (BIM) method was often used in the construction sector to make the construction work process efficient. Along with the increase in conservation and preservation activities of historical buildings, the BIM method has also begun to be applied to historic buildings, so the term Historical BIM (HBIM) method has emerged [2].

Various studies and practical activities have proven the HBIM method's reliability in documenting and scanning all information on metadata-related historical building objects, such as historical data, conservation policies, and significance values [3], [4]. The accurate and representative 3D modeling supports efforts to develop a digital management system to preserve and preserve historic buildings. Realistic 3D building models benefit simulation purposes [5], [6]. A digital survey technique is needed to produce detailed 3D models that are

attractive and useful for various purposes in cultural heritage [7]. Image-based virtual 3D modeling technology currently uses three approaches: sketch-based modeling, for example, SketchUp; procedural grammar-based modeling; and photogrammetric-based modeling [8].

The formation of three-dimensional media required expertise and skills in mastering special software. One method of forming 3D objects is by using the photogrammetry method. So many photos captured by the camera will have digital information, then put together into software and processed into 3D objects [9]. Making 3D objects with a high level of difficulty and detailed object shapes in the old way can take a long time, while the photogrammetry method can save time [10]. This method has an accurate result, an effective for wide-area using aerial photos and saving time and money. Implementation of modeling 3D objects using photogrammetry varies in many areas like Agricultural Land Modeling [11], Aerial mapping [12], and infrastructure site monitoring [13].

To help recognize the shape of the temple, modeling is carried out through 3D software by applying a workflow or photogrammetry method steps. Thus, it is necessary to create a standard workflow for making 3D models through the photogrammetry method because of the complexity of the process and workflow. So, this research aims to design the workflow for making the 3D object based on the photogrammetry approach into a formal modeling framework.

This paper proposes a standard framework for developing 3D objects of historical relics based on photogrammetry methods. This framework can be used for making other 3D objects which are still a lot of historical objects in Indonesia. The framework, workflow, and the results of the case on the development of the Badut Temple 3D object from this study provide an important role as a model that can be applied to other cases efficiently and effectively. This research is also a part of making Indonesia's cultural heritage digital preservice database [14]. The framework is important as a standard guideline for creating many historical relics in Indonesia.

II. FRAMEWORK FOR HISTORICAL RELICS MODELLING

The method used for making a 3D model is the photogrammetry method. Based on the experience in developing 3D models using the photogrammetry method, we propose a workflow framework for making 3D models of historical relics, as shown in Fig. 1. This framework consists of four consecutive processes that must be carried out

*Corresponding Author.

carefully and thoroughly. These steps are coded as follows: A) Preparation; B) Photo Capturing; C) Photo Processing & Modelling; D) Finalization.

The framework starts with the preparation phase by determining which object will be processed into a 3D model, then visiting the site to see the situation while preparing all the equipment. The types of equipment to be prepared are a high resolution of digital camera and an Unmanned Aerial Vehicle (UAV) or drone. The site visit is important to make sure that the process for taking object photos is permitted since the object is a cultural heritage. While visiting the location, it needs to take some initial photos to determine the photo capture flow in the stage of photo planning. A document template of object data is used for structuring the plan, process, and the results as shown in Fig. 2.

The photo capturing process should start by creating the photo capture flow, as shown in Fig. 3. Photo capture flow consists of the plan for the spot for taking photos circularly of the object. It consists of 2 stages of the photo plan, the first is around the object (yellow circle), and the second is circularly from the top of the object (blue circle). Then photos are taken using a digital camera from each position, including the drone's top position. To ensure that all photos are clear and bright enough, it is better to take them in the morning when the sunlight is clear enough. Since the number of digital photos is huge, it should be documented carefully.

After all of the photos have been captured, it is ready to proceed using special software that implements photogrammetry. Several softwares like Autodesk Recap, Regard3D, COLMAP, Meshroom, and RealityCapture can be candidates for this process. Table I shows the comparison between this software. RealityCapture is the best software but the most expensive, while Meshroom is good enough with a better price [15]. Meshroom software is chosen because it has excellent fidelity and can work better in low-light situations. So, all processes of making 3D models from a series of photos can be implemented. Meshroom will work according to the stages of modeling work.

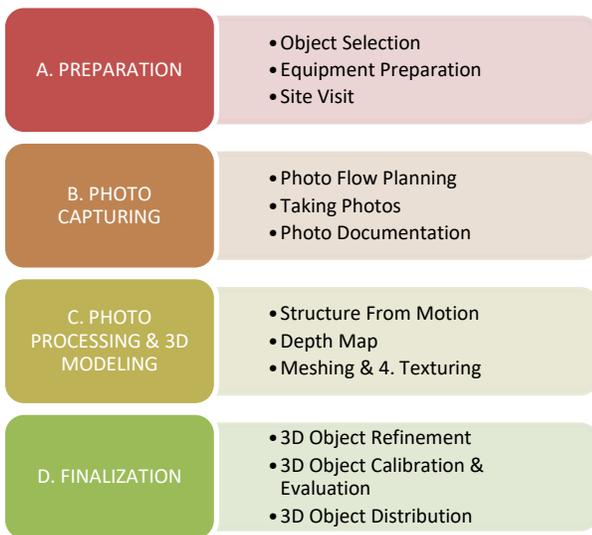


Fig. 1. Framework for Development of 3D Temple Object.

3D HISTORICAL OBJECT FRAMEWORK by MGM TECH LAB UB	
OBJECT DATA	
OBJECT ID:	Ej-Malang-01
Object Name:	Candi Badut
Type:	Temple
Location:	Lat: -7.939980
	Long: 112.632721
Address:	Jl. Raya Candi V No.5D, Doro, Karangwido, Dau
City:	Malang City
Province:	East Java
Web Data:	https://candi.nusantara360.id/web/
PHOTO CAPTURING	
Site Visit:	April 2021
Photo Capture:	May 2021
No of photos:	208
Directory:	3D Candi Badut / photos
Photographer:	Benjamin
Device:	Drone E88 Pro 4K Ultra HD Camera
3D OBJECT RESULTS	
File Name:	1. CandiBadut.dae 2. Texture0.001.jpg
Format:	dae (Collada)
File Size:	Texture: 211.902 KB Texture: 4.645 KB
Directory:	3D Candi Badut / results

Fig. 2. Framework Document of 3D Temple Object Database.

There are four processes for creating 3D models. It starts with making the structure from motion, which is the stage for obtaining a point cloud. After the point cloud is obtained, create a depth map using the DepthMap function. The DepthMap stage connects the point clouds to form a field. After the DepthMap stage is complete, the work process continues to the Meshing stage. At this stage, the fields that have been obtained begin to be put together to form a 3D model. Then the last stage in Meshroom is the Texturing stage. At this stage, the 3D object forms a texture that matches the original object in terms of color, shape, and position.

The last process is finalization, consisting of refinement, evaluation, and distribution. The 3D model from the previous process should be refined using another 3D tool. At this stage, software like Blender™ can be used to remove all unnecessary objects in the 3D model. The final 3D model is then evaluated by calculating the real size of the real temple object compared to the size of the 3D models. Some refinement is needed to resize the 3D model if there is a different size between the real and the temple model. The last step is to distribute the 3D model into a common 3D object file format that can be exchanged between software.

TABLE I. COMPARISON OF PHOTOGRAMMETRY SOFTWARE

Software	Regard3D	COLMAP	Meshroom	Reality Capture
Quality	***	****	****	*****
Speed	***	****	**	*****
Features	**	**	**	****
User Friendly	***	*	**	*****
Price	*****	*****	*****	**
Shape Fidelity	***	****	*****	*****
Details	***	****	***	*****
Noise	*	***	**	****
Low Lit Area	**	***	****	****
Overall	**	***	****	*****

III. PROCESS, RESULTS AND ANALYSIS

A. Object Selection and Preparation Process

The first thing to do in 3D modeling of historical relics is select the objects in the real world to be photographed into 3D objects. In this study, we chose the Badut Temple as the object to be studied. Initial site visiting has to be done to get some information and site photos to make the photo capture plan. The following is a view from the front side of the Badut Temple in Fig. 3. All process was planned in a document template as shown in Fig. 2.



Fig. 3. Front view of Badut Temple.

B. Photo Capture Flow Design

Before taking all the photographs of the temple, a photo capture flow has to be planned. The top view of the temple from Google Maps is used to draw the photo capture plan, as shown in Fig. 4. There are two stages to circularly taking photos; the first stage took 104 photographs with the distance of taking photographs (point A) to the object (point O) as far as 10 meters. The second stage took 104 photos with a photo-taking distance (point P) to the object (point O) as far as 3 meters. The angle between photographs is 3.5 degrees (A-O-B Angle and P-O-Q Angle).



Fig. 4. Badut Temple Photo Capture Flow.

C. The Results of Taking Photos of the Badut Temple

All photos were taken at the Badut Temple location using a high-quality digital camera and drone. As defined before, taking photographs of the Badut Temple was done in a circle, as shown in Fig. 4. The results of the Badut Temple photoshoot were obtained from as many as 208 photos and stored in a folder, as shown in Fig. 5. The number of photos varies depending on the size of the object.

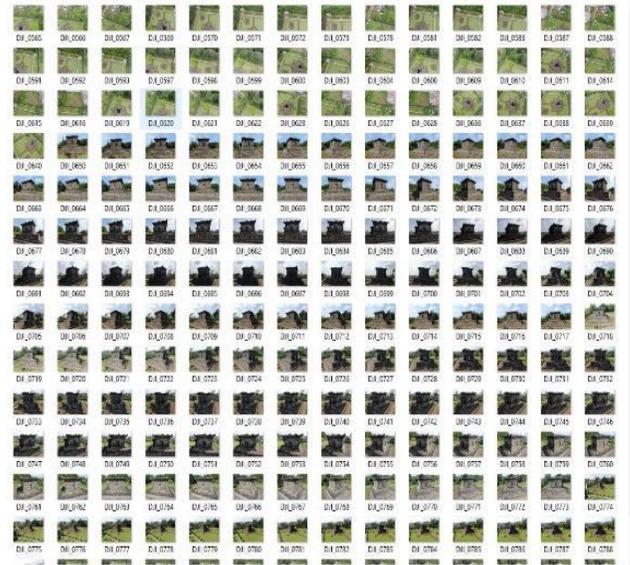


Fig. 5. The Results of Taking Photos from the Badut Temple.

D. Photos Processing in Meshroom

After all of the photos are collected, we can proceed with the processing steps. *Meshroom* is used in this phase and consists of five processing steps: importing photos, structuring, creating a depth map, meshing, and texturing.

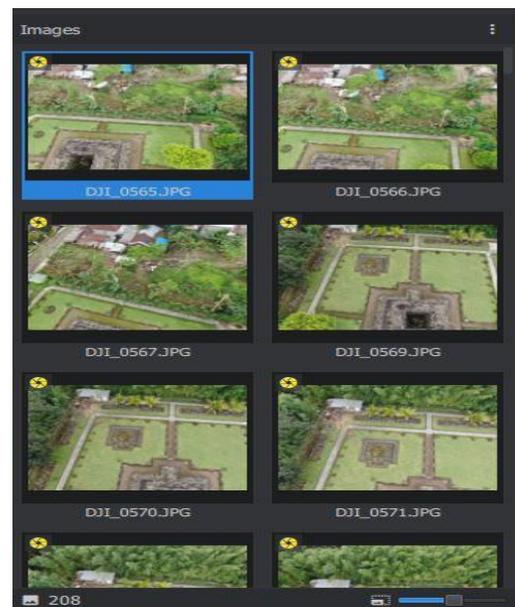


Fig. 6. Import Badut Temple Photos.

1) *Import stage of badut temple photos in meshroom:* Drag and drop photos saved in a folder into Panels Images on Meshroom. After the photograph enters the Panel Images, the work file is saved first. Then click start to start the operation. Display the Images panel in Meshroom as shown in Fig. 6.

2) *Stage structure from motion photo of badut temple in meshroom:* After the photos are put into the software, Meshroom will process them to generate the point cloud. The point cloud will be shown from the photo processing results. In Fig. 7, the point cloud obtained from this stage is 191,374 point clouds. This number already gives an idea of how the 3D model of the Badut Temple will be.

3) *Stage depthmap photo of badut temple in meshroom:* At this stage, the point cloud process obtained from the previous stage is more focused on unifying the point cloud so that it is close to the similarity to the object. The process at this DepthMap stage takes a long time and is prone to failure. Here's what the DepthMap stage looks as shown in Fig. 8.

4) *Meshing stage badut temple photo in meshroom:* At this stage, 3D objects from the Badut Temple have begun to be seen along with other objects caught by the camera. The shape of the Badut Temple 3D mesh is shown in Fig. 9.

5) *Texturing stage of badut temple photos in meshroom:* At this stage, the texturing process (shape and color) matches the texture of the real object of the Badut Temple. The 3D of the Badut Temple is shown in Fig. 10. This stage is the final stage of creating 3D objects in Meshroom software.

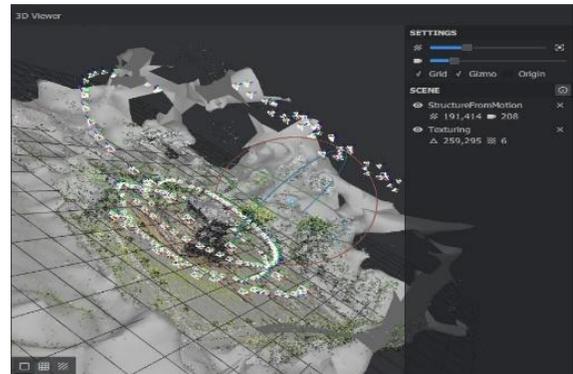


Fig. 9. Badut Temple Meshing Stage.



Fig. 10. Badut Temple Texturing Stage.

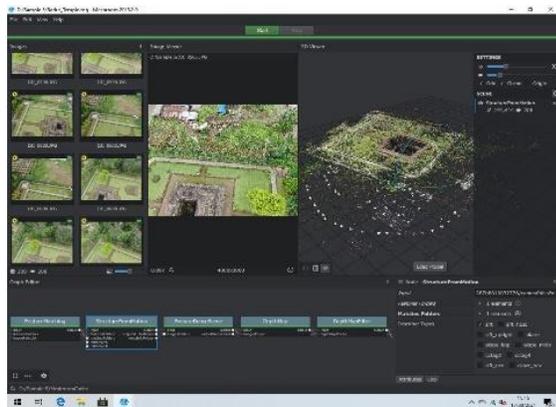


Fig. 7. The Stage Structure of the Badut Temple Movement.

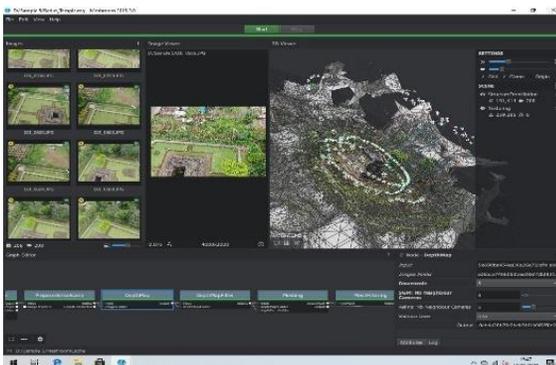


Fig. 8. Badut Temple DepthMap Stage.

After the Texturing stage is complete, the 3D object results in the Meshroom software are in the form of files that can be used for digital assets later. The following is a 3D file of the Badut Temple from the processing results in the Meshroom, as shown in Fig. 10.

In Fig. 11, there are several types of files from the processing results in Meshroom, which have different functions. The PNG file type results from the texture on the Badut Temple object, which contains a 2D image representation of the surface mapping of the Badut Temple 3D object. The MTL file type is a file that contains material information for the 3D Badut Temple object. A file of type 3D Object is a file that contains the final 3D object obtained. Files with this type of 3D Object will be used later for further work; in this case, the work is continued in Blender 3D.

Name	Date modified	Type	Size
log	17/03/2021 15:44	File	380 KB
statistics	17/03/2021 15:44	File	48 KB
status	17/03/2021 15:44	File	2 KB
texture_1001	17/03/2021 15:23	PNG File	31.544 KB
texture_1002	17/03/2021 15:27	PNG File	36.471 KB
texture_1003	17/03/2021 15:32	PNG File	46.115 KB
texture_1004	17/03/2021 15:36	PNG File	40.127 KB
texture_1005	17/03/2021 15:40	PNG File	42.095 KB
texture_1006	17/03/2021 15:44	PNG File	8.102 KB
texturedMesh.mtl	17/03/2021 15:19	MTL File	1 KB
texturedMesh	17/03/2021 15:19	3D Object	17.529 KB

Fig. 11. Badut Temple Model 3D File.

E. 3D Object Model Refinement

The file of type 3D Object from the Meshroom results is then imported into the Blender 3D software, as shown in Fig. 12. When taking photos, the temple object is always focused in the middle, so the result of importing the 3D model of the Badut Temple will make the temple object located in the middle, among objects caught on camera but unused.

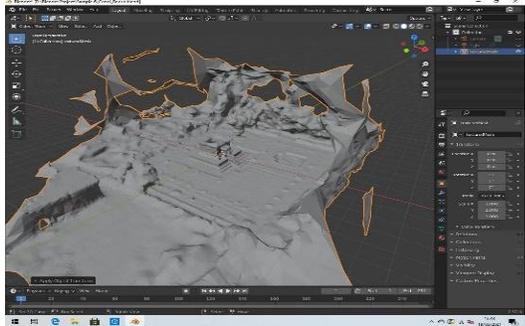


Fig. 12. Imported 3D Badut Temple Model.

At this stage, the removal or deletion of unused objects around the temple is carried out. The unused object is highlighted, and the deletion process is carried out. Then the final result of the 3D model in Blender 3D is obtained, as shown in Fig. 13.

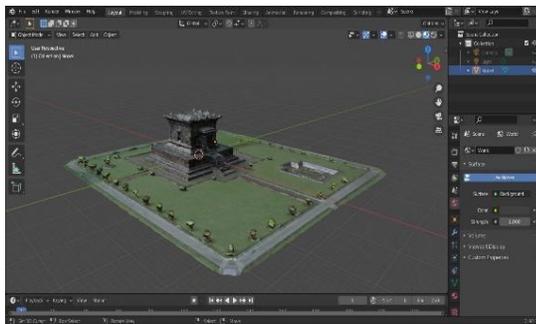


Fig. 13. 3D Results of the Badut Temple Model.

F. 3D Object Evaluation

To calculate the level of accuracy of the 3D model, the model size has to be compared with the temple's original size. A measuring tape was made to get the measurement results in the field. Meanwhile, to measure the size of 3D objects in Blender 3D software, use the measure function. Both results were then compared to ensure that the 3D model has a proportional size to the model of a real object in a small size. This is also for calibration purposes.

1) *Field measurement results:* The results of measurements in the field are taken by measuring the parts of the temple. In this case, 15 samples were taken from each temple part using a measuring tape. For the sample of the front view of the temple, ten samples of the length dimensions D1 to D10 were taken, as shown in Fig. 14. For the left view sample of the temple, five samples of the length dimensions D11 to D15 were taken, as shown in Fig. 15. Then the measurement results are obtained as in Table II.

TABLE II. FIELD MEASUREMENT RESULTS

Dimension	Field Results (m)
D1	4.212
D2	1.509
D3	4.312
D4	1.243
D5	2.274
D6	2.205
D7	1.994
D8	1.300
D9	1.333
D10	2.265
D11	10.807
D12	7.456
D13	6.138
D14	1.565
D15	1.308



Fig. 14. The Size of the Front View of the Badut Temple.



Fig. 15. Size of the Badut Temple's Left View.

2) *3D Model measurement results:* To measure the size of 3D objects in Blender 3D software, use the measure function. Measurements on the software, adjusted to the sample on measurements in the field. For the sample of the front view of the temple in 3D Blender, ten samples were taken with dimensions D1 to D10, as shown in Fig. 16. For the left view sample of the Temple in 3D Blender, five samples were taken with dimensions D11 to D15, as shown in Fig. 17. All the results are obtained in Table III.

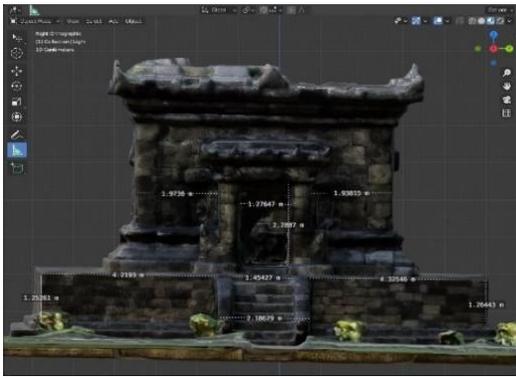


Fig. 16. Front View Size in Blender 3D.



Fig. 17. Left View Size in Blender 3D.

TABLE III. MEASUREMENT RESULTS IN BLENDER 3D

Dimension	3D Blender Result (m)
D1	4.2193
D2	1.4542
D3	4.3254
D4	1.2764
D5	2.2807
D6	1.9736
D7	1.9381
D8	1.2526
D9	1.2644
D10	2.1862
D11	10.802
D12	7.4809
D13	6.1213
D14	1.5130
D15	1.2526

3) *RMSE calculation results:* After measuring the size of the real object and 3D model, the *Root Mean Square Error* (RMSE) formula is used to measure the differences between them using the equation (1).

$$RME = \sqrt{\frac{\sum (Y_i - \hat{Y})^2}{n}} \quad (1)$$

The first thing to do is to calculate the sum of the squares of the differences between the data obtained from 3D models and the data of the real object, as shown in Table IV. The summary obtained from Table IV is entered into the formula for calculating RMSE.

$$RMSE = \sqrt{\frac{0.080895 \text{ m}^2}{15}}$$

$$RMSE = 0.073437 \text{ m}$$

$$RMSE \approx 7.34 \text{ cm}$$

The accuracy of the 3D model of the Badut Temple obtained is 0.0734 meters, or equivalent to 7.34 cm. From the calculation results, the results obtained are smaller than 10 centimeters, so it can be said that the 3D Badut Temple model generated from the photogrammetry method is close to the shape of the Badut Temple in the field (original). Comparison of the Badut Temple photos and the processing of the Badut Temple in Meshroom and Blender 3D software as shown in Fig. 18 to 21.

4) *3D Object distribution:* The results obtained from Blender 3D are 3D objects in various files format, one of which is the Collada (.dae) file. The visualization of the Blender 3D results was obtained as shown in Fig. 22. The files obtained from Blender 3D can be used in other 3D processing software. The 3D model of the Badut Temple can be used as a digital asset for many purposes like digital preservation, video games, video animations, and others.

TABLE IV. MEASUREMENT RESULTS IN BLENDER 3D

Dimension	3D Model (Y _i) (m)	Real Object (Ŷ) (m)	(Y _i - Ŷ) ² (m ²)
D1	4.2193	4.212	0.000053
D2	1.4542	1.509	0.002995
D3	4.3254	4.312	0.000181
D4	1.2764	1.243	0.001120
D5	2.2807	2.274	0.000045
D6	1.9736	2.205	0.053546
D7	1.9381	1.994	0.003119
D8	1.2526	1.300	0.002246
D9	1.2644	1.333	0.004702
D10	2.1862	2.265	0.006195
D11	10.802	10.807	0.000018
D12	7.4809	7.456	0.000624
D13	6.1213	6.138	0.000277
D14	1.5130	1.565	0.002704
D15	1.2526	1.308	0.003068
Amount (m ²)			0.080895



Fig. 18. Comparison of Front View Photos and 3D Badut Temple.



Fig. 19. Comparison of Left View Photos and 3D Badut Temple.



Fig. 20. Comparison of Rear-View Photos and 3D Badut Temple.



Fig. 21. Comparison of Right View Photo and 3D Badut Temple.



Fig. 22. Visualization of the Badut Temple.

IV. CONCLUSION

This paper proposed a framework for digitizing historical relics using a photogrammetry approach. The process consists of 4 phases: Preparation, Photo capturing, Photo Processing & Modelling, and Finalization. All these processes can complete in just one week. Implementation of this framework in Badut Temple shows that the photogrammetry approach is efficient for creating a 3D model of historical relics.

In the photogrammetry approach, the series of circular views of the object photo is taken using a high-resolution

camera and UAV drone. Then all the photos are processed carefully using Meshroom software in four steps: structuring, depth map, meshing and texturing. Special software for processing 3D objects like Blender is then used for refinement and finalization. The result of the accuracy of the 3D model of the Badut Temple is below 10 centimeters. It means that the model has a precise size compared to the real object. So, this approach is also effective for creating 3D models precisely while maintaining the detail of the reliefs.

V. FUTURE WORK

The 3D model of the Badut Temple can be used as a digital asset for many purposes like digital preservation, video games, virtual reality, and others. This model will then use as part of the digitalization and creation of the digital database of historical relics in Indonesia. Our research group of Media, Game, and Mobile (MGM) Technology at the Computer Science Faculty of Brawijaya University is researching Indonesia's digital heritage of historical buildings, starting from many of the relics in East Java, Indonesia. We develop a web portal and digital database of the model of historical buildings in Indonesia, especially the temple, as shown in Fig. 23.

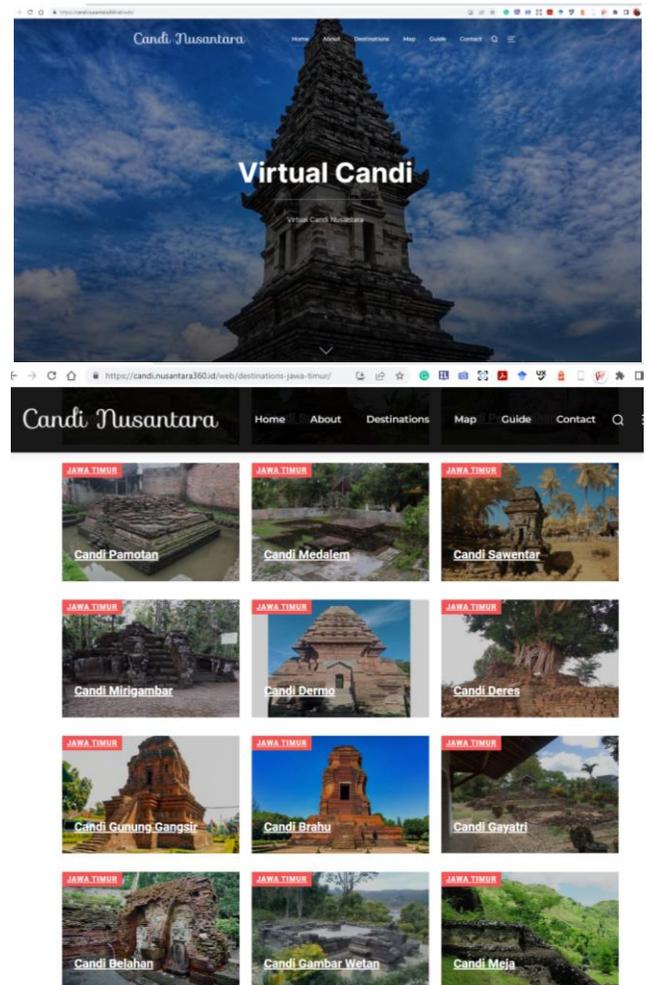


Fig. 23. Portal of Virtual Candi as Digital Database of Indonesian Historical Temple.

Since the temple wall has many reliefs that contain a visualization of a story, the 3D model created from this approach should be able to preserve it. Future work must be done to ensure that all the reliefs still exist. Digital restoration and refinement processes may proceed to make the reliefs exist like the original relics.

ACKNOWLEDGMENT

We thank all our participants for their contributions. This research work is funded by the Faculty of Computer Science at Brawijaya University and is part of making Indonesia's cultural heritage digital preserve database.

REFERENCES

- [1] Yuwono, Pratomo, D., G., Mulyono, Y., E., R., 2018. Rekonstruksi Model 3D Candi Jawi Dengan Metode Structure From Motion (SfM) Foto Udara. ITN Malang.
- [2] Santosa H, Yudono A, Adhitama M S . The digital management system of the tangible culture heritage for enhancing historic building governance in Malang, Indonesia[J]. IOP Conference Series Earth and Environmental Science, 2021, 738(1):06.
- [3] Jordan-Palomar, I.; Tzortzopoulos, P.; García-Valdecabres, J.; Pellicer, E. Protocol to Manage Heritage-Building Interventions Using Heritage Building Information Modelling (HBIM). *Sustainability* 2018, *10*, 908. <https://doi.org/10.3390/su10040908>.
- [4] Rocha, G.; Mateus, L.; Fernández, J.; Ferreira, V. A Scan-to-BIM Methodology Applied to Heritage Buildings. *Heritage* 2020, *3*, 47-67. <https://doi.org/10.3390/heritage3010004>.
- [5] Hullo J F, Grussenmeyer P and Fares S Photogrammetry and dense stereo matching approach applied to the documentation of the cultural heritage site of Kilwa (Saudi Arabia) 22nd CIPASymp. Kyoto Japan 2009 (Japan) p. 1–6.
- [6] Grussenmeyer P, Alby E, Assali P, Poitevin V, Hullo J F and Smigiel E Accurate documentation in cultural heritage by merging TLS and high-resolution photogrammetric data Proc. of SPIE - The International Society for Optical Engineering 2011 vol. 8085 <https://doi.org/10.1117/12.890087>.
- [7] Chiabrando F, Donadio E, and Rinaudo F SfM for orthophoto to generation: a winning approach for cultural heritage knowledge ISPRS - Intl. Arch. of the Photogram., Rem. Sens. and Spat. Inf. Sci. 2015 XL-5/W7 <https://doi.org/10.5194/isprsarchives-XL-5-W7-91-2015>.
- [8] Singh S P, Jain K, and Mandla V R A new approach towards image-based virtual 3D city modeling by using close-range photogrammetry ISPRS Ann. of Photogram., Rem. Sens. And Spat. Inf. Sci. 2014 II-5 pp 329–37 <https://doi.org/10.5194/isprsannals-II-5-329-2014>.
- [9] Sondang, V.A., 2017. Pembuatan Model Ortofoto Hasil Perkaman dengan Wahana UAV Menggunakan Perangkat Lunak Fotogrametri.
- [10] Lachambre, S., Lagarde, S., Jover, C., 2017. *Photogrammetry Workflow*. Unity 3D.
- [11] Faisal Mahmood, Khizar Abbas, Asif Raza, Muhammad Awais Khan and Prince Waqas Khan, "Three Dimensional Agricultural Land Modeling using Unmanned Aerial System (UAS)" International Journal of Advanced Computer Science and Applications (IJACSA), 10(1), 2019.
- [12] Muhammad Yazid Abu Sari, Abd Wahid Rasib, Hamzah Mohd Ali, Abdul Razak Mohd Yusoff, Muhammad Imzan Hassan, Khairulnizam M.Idris, Asmala Ahmad and Rozilawati Dollah, "3D Mapping based-on Integration of UAV Platform and Ground Surveying" International Journal of Advanced Computer Science and Applications (IJACSA), 9(12), 2018.
- [13] Cristian Benjamin Garcia Casierra, Carlos Gustavo Calle Sanchez, Javier Ferney Castillo Garcia, and Felipe Munoz La Rivera, "Methodology for Infrastructure Site Monitoring using Unmanned Aerial Vehicles (UAVs)" International Journal of Advanced Computer Science and Applications (IJACSA), 13(3), 2022.
- [14] Yudono, A., Santosa, H., & Tolle, H. (2020, July). The Three Dimension (3D) Spatial Urban Heritage Informatics of Malang City, Indonesia. In International International Conference of Heritage & Culture in Integrated Rural-Urban Context (HUNIAN 2019) (pp. 124-129). Atlantis Press.
- [15] Photogrammetry: Step-by-Step Guide and Software Comparison. <https://formlabs.com/asia/blog/photogrammetry-guide-and-software-comparison/> [Accessed: April 2022].

Dual-fast Greedy Heuristic Algorithm for Green ICT

Inas Abuqaddom

Dept. Computer Science

King Abdullah II School of Information Technology
The University of Jordan, Amman, Jordan

Abstract—The Internet power consumption represents 3.6% to 6.2% of the annual worldwide power consumption and is continually expanding. The awareness of this problem has increased, hence, a few strategies are being put into place to decrease the power consumption of the Information Communication Technology (ICT) sectors, in general. Backbone networks are the main part of the Internet power consumption because their line cards expend a lot of energy, also their links are commonly bundled and provide larger capacity than needed. Therefore, bundled links are partially shut down during times of low demand to reduce power consumption. Literature introduces a few heuristic algorithms that are run periodically to shut down bundled links partially. This paper proposes a Dual-Fast Greedy Heuristic algorithm (DGH), which significantly speeds up the power savings. DGH is examined on the topology and traffic of the Abilene backbone. The experimental results show that DGH provides competitive power savings with minimum execution time.

Keywords—Abilene backbone; backbone network; bundled link; capacity; Internet power consumption; power savings

I. INTRODUCTION

ICT (Information Communication Technology) handles the processes of communications such as telecommunications, broadcast media, intelligent building, network-based control, etc. [1, 2]. Green ICT is responsible for using computing resources efficiently and effectively with minimum impact on the environment, mainly by reducing their power consumption [2, 3]. Since the Internet is the pivot sector of ICT, network researches focus on reducing power consumption. The most popular power-saving technique is based on the power consumption of servers and wireless equipment [4]. However, reducing the power consumption of wired networks has been ignored, even though it is critical [5]. For example, powering the wired networks in the United States alone expenses an expected 0.5-2.4 billion dollars per year. Additionally, network architectures having better energy efficiency allow deploying networks in poor infrastructures [6].

The Internet has multiple backbone networks. Since a backbone network interconnects networks and provides paths for data exchange. Also, a backbone network is called a core network. Usually, the capacity of a backbone link is larger than the needs of backbone networks. The additional capacity is used to cover traffic shifts and to provide alternative paths for broken links [7]. For example, the average used capacity in backbone networks of big Internet service providers is no more than 30-40%, consequently, there are 70-60% extra capacities. Accordingly, using dynamic capacity instead of static capacity

for backbone networks will reduce the power consumption efficiently. The optimal technique to provide dynamic capacity is that the backbone links are partially shut down and powered as needed. Since the used capacity through off-peak hours is reduced to one-third or more of peak hours [5, 8].

A backbone link connects two routers and is structured as multiple physical cables that are dealt with as one logical bundled link [9]. Generally, a logical bundled link with all its physical cables is called bundled link, aggregate link, or composite link [9, 10]. Additionally, there is a line card at each end per physical cable to serve it. Nonetheless, bundled links are standardized by IEEE 802.1AX [10]. The capacity of a bundled link is the aggregate capacities of all its physical cables. Thus, to upgrade a bundled link i.e. increasing its capacity, you just add more physical cables to the existing cables. As a result, the capacity of a bundled link may exceed the capacity of the fastest physical cable. For example, given a bundled link of five OC-192 cables each with a 10 Gbps capacity, then the bundled link capacity is 50 Gbps.

The optimal power-saving approach shuts down and powers some physical cables of a bundled link as needed. In other words, this approach is an optimization problem that maximizes the power savings by shutting down the most possible physical cables of every bundled link, as it yet has enough capacity for future traffic. The physical cable selection is based on the current and expected traffic matrix, the network topology, and the bundled link capacity [11]. Thus, heuristic algorithms are used to find the optimal selection, which is an NP-complete problem [12]. Accordingly, these algorithms extremely vary in execution time, which increases the router overhead. However, the execution time of these algorithms is based on how they select physical cables of a bundled link to be shut down or powered [11].

This paper proposes the Dual-fast Greedy Heuristic (DGH) algorithm to reduce the power consumption of backbone networks with limited overhead on routers. DGH shuts down some physical cables and their corresponding line cards. Since line cards consume most of the router power consumption [8]. Moreover, DGH provides competitive power savings compared with other algorithms, in addition to its simplicity and high speed.

The rest of this paper is organized as follows; Section II describes the problem notations. Section III analyzes the literature review. Section IV portrays the proposed algorithm. Section V shows the experimental results. Finally, conclusions are shown in Section VI.

II. PROBLEM NOTATIONS

The backbone network topology is described as a directed graph $G(V, E)$ as V is a set of routers and E is a set of links. Usually, links are bundled and every bundled link $(u, v) \in E$, such that (u, v) connects two routers; $u, v \in V$ and has a capacity $c(u, v)$. Every bundled link consists of B physical cables. For example, a bundled link (u, v) consists of five 10 Gbps physical cables. Then $B=5$, which is the bundle size, and $c(u, v)=50$ Gbps, which is the bundled link capacity. The demand i.e. traffic between a couple of routers is described as a row (s_d, t_d, h_d) in the traffic matrix D , where s_d is the source router, t_d is the destination router, and h_d is the amount of demand between s_d and t_d combination. Furthermore, let $f_d(u, v)$ be the flow of the bundled link (u, v) and d is a demand amount through a bundled link (u, v) [13]. The aggregate flow of a bundled link (u, v) is denoted as $f(u, v)$ and shown in Equation (1). Usually, the flow $f(u, v)$ of a bundled link (u, v) does not exceed its capacity $c(u, v)$ [14].

$$f(u, v) = \sum_D f_d(u, v) \quad (1)$$

Furthermore, the extra capacity concept is the aggregate capacity of all unused physical cables of a bundled link. For example, Fig. 1 shows a partial backbone network, assuming there is a demand $d=4.5$ Gbps between a source router $s_{4.5}$ and a destination router $t_{4.5}$. Assume that every bundled link has a capacity of 10 Gbps, which are the aggregate capacities of ten 1 Gbps physical cables. As shown in Fig. 1, there are two paths between s and t , either $(s, 1, 2, 3, 4, t)$ or $(s, 4, t)$. Even though the demand is the same for both paths, its corresponding total flow varies. The total flow through the $(s, 1, 2, 3, 4, t)$ path is 22.5 Gbps, so the total extra capacities among all bundled links are 37.5 Gbps. However, the total flow through $(s, 4, t)$ path is 9 Gbps and the total extra capacities among all bundled links are 51 Gbps. Therefore, to minimize the total flow and maximize the total extra capacities, you have to select the shortest paths.

Accordingly, to maximize the total extra capacities of all bundled links, the traffic is routed through the possible shortest paths. Thus, the total flow through all bundled links is reduced. In other words, routing traffic through shortest paths minimizes the total flow of all bundled links and maximizes the total extra capacities of all bundled links, as shown in Equation (2).

$$\min \sum_{(u, v) \in E} f(u, v) \quad (2)$$

The network-management system runs an optimization algorithm periodically to reduce power consumption. The inputs of an optimization algorithm are a network topology $G(V, E)$, a bundle size B , and a traffic matrix D . Then, the optimization algorithm defines a network setup that utilizes the least physical cables from all bundled links, with the end goal of fulfilling all demands. The number of powered physical cables in a bundled link (u, v) is denoted as $n_{u, v}$. In other words, the outputs of an optimization algorithm are the selected powered physical cables per bundled link and the rerouting paths that some demands may use to increase the utilization of powered physical cables in the network [14].

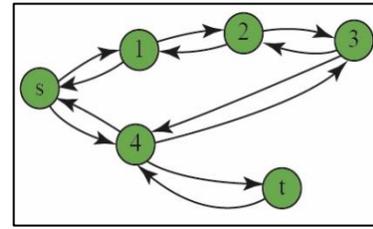


Fig. 1. Backbone Network using Hierarchical Topology.

III. LITERATURE REVIEW

To control power consumption in local-area networks, Ethernet cards utilize off-peak periods to work in low power consumption modes. For example, Broadcom and Intel produce network cards containing programmable rest clocks, that can be controlled by algorithms to shut down a link for a while [15]. Nonetheless, some approaches shut down a router totally including all its bundled links and their corresponding line cards at that router [11]. However, these approaches are not compatible with wide area networks, because shutting down the whole bundled link must produce a packet loss [15].

In wide-area networks, literature shows approaches to control power consumption. One approach uses a sleep mode, such that the system puts network cards into the sleep mode. However, how does the system deal with traffic through sleeping network cards? There are two solutions; either using coordinated sleeping routers or uncoordinated sleeping routers. The coordinated sleeping router is a centralized solution, that reroutes traffic through sleeping network cards into alternative active network cards as could reasonably be expected. The drawback of this solution is that the router needs a dynamic protocol [6, 16, 17].

The uncoordinated sleeping router is an uncentralized solution. Every network card notifies its neighboring network cards before going to the sleep mode, which is allowable at low demand periods. Thus, an active network card wakes up the sleeping network card as needed by sending a wake-up packet. The drawback of this solution is the latency because of wake-up time and neighboring processing [6]. Furthermore, the authors of [15, 19] recommend an extreme suggestion as all network equipment should support a slow-speed mode to reduce power consumption.

Nevertheless, literature shows that the optimal power-saving approach shuts down and powers some physical cables of a bundled link according to the low demand time. This approach uses the extra capacities, which are determined by the traffic matrix of all bundled links. Generally, this approach maximizes the total extra capacities of all bundled links since the traffic is routed through the possible shortest paths. As mentioned in the Introduction Section, this approach uses heuristic algorithms to find a possible optimal selection among physical cables of a bundled link [6]. This section presents the most known heuristic algorithms, which utilize the unused capacity of powered physical cables. These algorithms are the Fast Greedy Heuristic algorithm (FGH), Exhaustive Greedy Heuristic algorithm (EGH), and Bi-level Greedy Heuristic algorithm (BGH) [19].

A. Fast Greedy Heuristic Algorithm

FGH is the fastest and simplest algorithm compared with EGH and BGH. Initially, FGH minimizes the total flow of all bundled links to maximize the total extra capacities of all bundled links, as shown in (2). Then FGH shuts down all extra capacities such that the remaining powered physical cables can serve all the network traffic. After that, FGH finds the physical cable with the largest unused capacity by using (3) [14].

$$f(u,v) \leq (n_{uv} \div B) c(u,v) \quad \forall (u,v) \in E$$

such that,

$$\max_{(u,v)} ((n_{uv} c(u,v) \div B) - f(u,v)) \quad (3)$$

Periodically, FGH attempts to shut down the physical cable having the largest unused capacity and reroutes its corresponding flow. According to the example in Section II, if you shut down one out of the remaining five physical cables, that carry out the demand $d=4.5$ Gbps, you have to reroute a demand $d=0.5$ Gbps into alternative paths. Then, FGH examines (2). If it is not satisfied anymore, FGH powers the shutting down physical cable and marks its bundled link (u, v) as “final”, so no more future attempts to shut down any of its physical cables. As long as (2) is satisfied, the shutting down physical cable is confirmed and (3) is calculated. FGH repeats until all bundled links are marked as “final” [14].

However, FGH has drawbacks; such as if FGH shuts down a physical cable that produces a suboptimal solution, it will never backtrack to revise the selection.

B. Exhaustive Greedy Heuristic Algorithm

EGH performs as FGH using different conditions to shut down a physical cable. EGH calculates a penalty value for every candidate-physical cable. Such that, EGH shuts down the physical cable having the smallest penalty. Nevertheless, the penalty for a physical cable is calculated based on the flow distribution before and after shutting down the physical cable. Thus, the penalty allows the algorithm to do a “look-ahead” decision on each physical cable before removing it. Usually, EGH finds an optimal selection consuming a larger execution time and increasing the router overhead because of penalty calculations [14].

C. Bi-level Greedy Heuristic Approach

BGH performs as EGH using the penalty condition, but in a different manner. BGH applies a penalty on a pair of physical cables and shuts down the pair having the smallest penalty. However, the penalty for a pair of physical cables is calculated based on the flow distribution before and after shutting down the physical cables. Therefore, BGH finds the optimal selection, consuming unreasonable execution time to make a removal decision [14]. Accordingly, the router overhead extremely increases because of the double penalty calculations per selection.

IV. PROPOSED ALGORITHM

As we mentioned previously, FGH is fast and simple. Literature shows that the three heuristic algorithms (FGH, EGH, and BGH) are close to each other in terms of power-saving amount, but they extremely vary in execution time.

Moreover, a network operator runs one of these algorithms very often to control the power consumption. Thus, it is essential to reduce the execution time, which is varied from a few minutes using FGH, a few hours using EGH, and countless times using BGH. Their execution times vary because of the (2) complexity, which is $O(|E|^2)$ for both FGH and EGH as is $O(|E|^3)$ for BGH. On the other hand, FGH and EGH are different in selecting a physical cable to be shut down. FGH selects a physical cable having the maximum unused capacity, as EGH selects a physical cable having the minimum penalty. However, penalty calculation is harder and consumes more time than unused capacity calculation. As a result, execution times of FGH and EGH are different, even though they have the same complexity of (2) [14].

Accordingly, this paper proposes a Dual-Fast Greedy Heuristic algorithm (DGH) to speed up the power-saving process consuming limited overhead on routers. Initially, DGH minimizes the total flow of all bundled links to maximize the total extra capacities of all bundled links. Then, DGH shuts down all extra capacities from all bundled links. After that, DGH randomly shuts down a physical cable from a bundled link (u,v) and reroutes the flow of the shutting-down physical cable into alternative shortest paths. As a test for optimality, DGH examines (2). If it is not satisfied, DGH powers the shutting down physical cable and marks its bundled link (u,v) as “final” to prevent future shut-down attempts on (u,v) . Otherwise, the shutting down physical cable is confirmed. Moreover, DGH repeats selecting a random physical cable from a random unmarked bundled link until all bundled links are marked as “final”.

DGH reduces the power-saving cost to the minimum because there is no calculation per selecting a physical cable. On the other hand, DGH could result in a suboptimal solution, since DGH randomly selects a physical cable. In other words, DGH is similar to other heuristic algorithms in the first part, which is shutting down all extra capacities from all bundled links. Then DGH randomly selects a physical cable to shut down without any extra calculations. Consequently, DGH provides power savings similar to FGH, EGH, and BGH, because of the first common part. Additionally, DGH consumes a lower execution time, because all further shutting down physical cables, after the first common part, are selected randomly.

V. EXPERIMENTAL RESULTS

The proposed algorithm DGH is examined using AMPL/CPLEX solver. CPLEX is an optimization package for linear, network, and integer programming. AMPL is an algebraic modeling language, which stands for A Modeling Language for Mathematical Programming. Generally, AMPL utilizes an optimization package such as CPLEX to solve optimization problems [20, 21]. As an experimental backbone network, the Abilene backbone network of 39 nodes is used. The experimental Abilene is examined using two topologies; Waxman and hierarchical topologies, as Table I presents their parameters, which are the number of bundled links as shown in “#Bundled links” column and the requested demands between any two nodes as shown in “Demands” column. The key difference between both topologies is connectivity. Such that

every existing bundled link between any two nodes is doubled to be in both directions for the hierarchical topology. However, some existing bundled links are doubled to be in both directions for the Waxman topology. Fig. 2 and 3 show examples of Waxman and hierarchical topologies, respectively. Results were collected on the Intel Core 2 processor running Ubuntu server 14.04.

TABLE I. PARAMETERS OF ABILENE TOPOLOGIES

Topology	#Bundled links	Demands
Hierarchical	148	2.450
Waxman	169	2.450

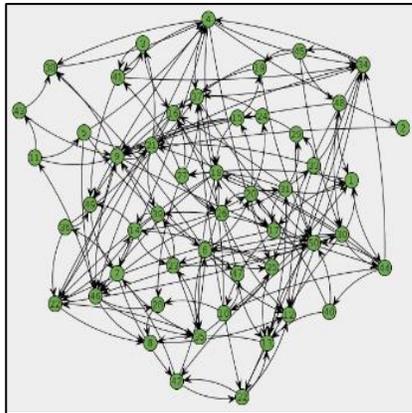


Fig. 2. Waxman Topology.

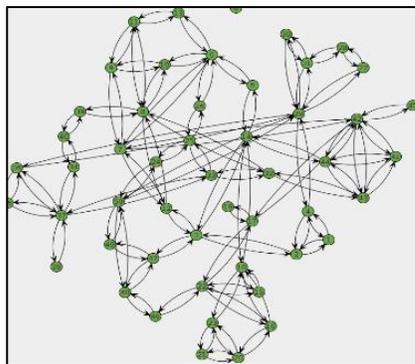


Fig. 3. Hierarchy Topology.

In the Waxman topology, the likelihood that two nodes are directly connected by a bundled link increases as the distance between them decreases. The hierarchical topology was created by GT-ITM [18]. Moreover, real demands were estimated by traditional entropy for urban traffic [22]. However, DGH is compared with FGH because it is the fastest algorithm.

Experimental results show that FGH and DGH provide similar power savings. Because both of them shut down all extra capacities from the beginning, also both of them fell into suboptimal solutions during the further steps. However, DGH outperforms FGH in terms of execution time. Fig. 4 shows the execution time of both FGH and DGH on Waxman topology. DGH outperforms FGH irregularly because of irregular topology. While Fig. 5 shows the execution time of both FGH and DGH on hierarchical topology. The curves of Fig. 5 reflect

the regular and little improvement of DGH due to the regular topology.

To translate the shown improvement into numbers, the improvement ratio is calculated using Equation (4):

$$IR = \text{avg}(T_{FGH}) - \text{avg}(T_{DGH}) \tag{4}$$

Where IR is the improvement ratio, as T_{FGH} and T_{DGH} are the execution time using FGH and DGH algorithms, respectively. Thus, DGH outperforms FGH in both topologies with improvement ratios of 18% and 36.17% on hierarchical and Waxman topologies, respectively.

Moreover, both Fig. 4 and 5 show that the bundled link size and the execution time are almost independent, because both DGH and FGH shut down all extra capacities in the first step.

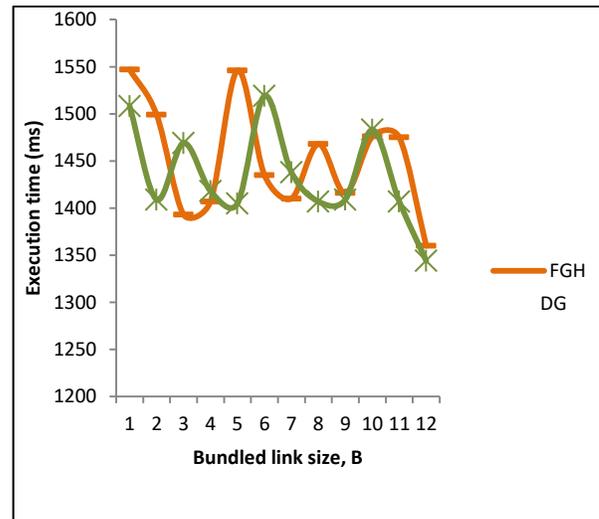


Fig. 4. The Execution Time of FGH and DGH Algorithms Applied on Waxman Topology.

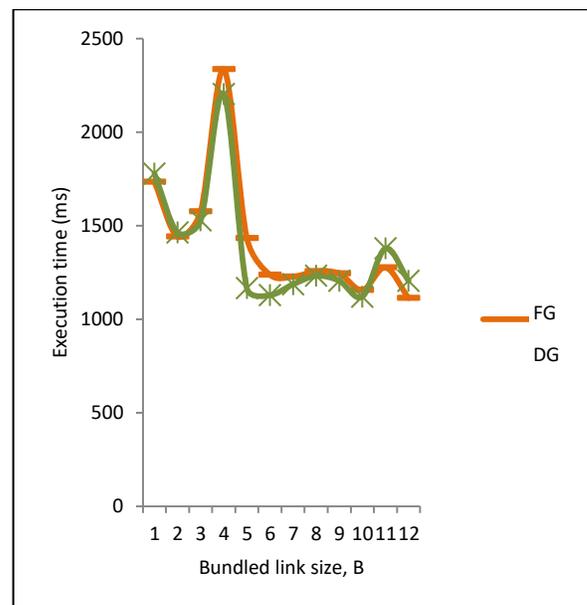


Fig. 5. The Execution Time of FGH and DGH Algorithms is Applied to Hierarchical Topologies.

VI. CONCLUSION

Green ICT is responsible for reducing the power consumption of computing resources to minimize their impact on the environment. The Internet represents up to 10% of the worldwide power consumption and is continually expanding. Furthermore, backbone networks are the main part of Internet power consumption. Since links of these networks are commonly bundled and provide larger capacity than needed.

A few approaches are proposed to reduce power consumption in the Internet backbone. One of them depends on shutting down individual cables of bundled links during times of low demand. However, optimal shutting down physical cables is an NP problem. Therefore, algorithms compete for increasing shutting-down cables in a reasonable time to provide more power savings. Accordingly, this paper proposes a dual-fast greedy heuristic algorithm (DGH), which shuts down all extra capacities from all bundled links. Then, DGH randomly shuts down a physical cable having an unused capacity. Also, DGH reroutes the flow of the shutting-down physical cable.

To assess DGH, the AMPL/CPLEX solver is utilized on the Abilene backbone. DGH is compared with the fastest algorithm, which is FGH. The experimental results show that DGH is faster than FGH and provides similar power savings as FGH. Nonetheless, the improvement ratios in terms of execution time are between 18% and 36.17% for Waxman and hierarchical topologies, respectively. The drawback of DGH is the suboptimal solution, which does not affect the power savings because DGH shuts down all extra capacities from the beginning. In a conclusion, DGH gets suboptimal selection to reduce the router overhead, which in turn reduces the power consumption. For future works, DGH could be examined using various backbone networks and various sets of parameters.

REFERENCES

- [1] A. Ozturk, et al., "Green ICT (information and communication technologies): a review of academic and practitioner perspectives,321" International Journal of eBusiness and eGovernment Studies vol. 3, no. 1, pp. 1-16, 2011.
- [2] MK. Dash, C. Singh, G. Panda, and D. Sharma, "ICT for sustainability and socio-economic development in fishery: a bibliometric analysis and future research agenda," Environment, Development and Sustainability, pp. 1-33, 2022.
- [3] JL. Hu, YC. Chen, and YP. Yang, "The development and issues of energy-ICT: a review of literature with economic and managerial viewpoints," Energies vol. 15, no. 2, p. 594, 2022.
- [4] Y. Cui, X. Ma, H. Wang, I. Stojmenovic, and J. Liu , "A survey of energy efficient wireless transmission and modeling in mobile cloud computing," Mobile Networks and Applications, vol. 18, no. 1, pp. 148-155, Feb 2013.
- [5] L. Chiaraviglio, M. Mellia, and F. Neri, "Minimizing ISP network energy cost: formulation and solutions," IEEE/ACM TRANSACTIONS ON NETWORKING, vol. 20, no. 2, APRIL 2012.
- [6] M. Gupta and S. Singh, "Greening of the Internet," Proceedings of ACM SIGCOMM, 2003.
- [7] JI. Castillo-Velázquez, and A. Delgado-Villegas, "GNS3 limitations when emulating connectivity and management for backbone networks: a case study of CANARIE," 2020 IEEE Canadian Conference on Electrical and Computer Engineering (CCECE), IEEE, 2020.
- [8] J. Chabarek, et. al., "Power awareness in network design and routing," IEEE INFOCOM, 2008.
- [9] RD. Doverspike, KK. Ramakrishnan, and C. Chase, "Structural overview of ISP networks," in Guide to Reliable Internet Services and Applications (C.Kalmanek, S. Misra, and Yang, eds.), Springer, 2010.
- [10] IEEE Computer Society, "IEEE standard 802.1AX: link aggregation," 2008.
- [11] L. Chiaraviglio, M. Mellia, and F. Neri, "Reducing power consumption in backbone networks," IEEE ICC, 2009.
- [12] L. Cui, et. al., "Joint optimization of energy consumption and latency in mobile edge computing for Internet of Things," IEEE Internet of Things Journal vol. 6, no. 3, pp. 4791-4803, 2018.
- [13] OM. Surakhi, M. Qatawneh, and HA. Ofeishat, "A parallel genetic algorithm for maximum flow problem," International Journal of Advanced Computer Science and Applications. vol. 8, no. 6, pp.159-164, Jan 2017.
- [14] W. Fisher, M. Suchara and J. Rexford, "Greening backbone networks: reducing energy consumption by shutting off cables in bundled links," ProceedingGreenNetworking '10 Proceedings of the first ACM SIGCOMM workshop on Green networking, pp. 29-34, 2010.
- [15] B. Heller, et.al., "Elastic tree: saving energy in data center networks," USENIX NSDI, 2010.
- [16] M. Islam, and M. Rashid, "A comprehensive analysis of blockchain-based cryptocurrency mining impact on energy consumption," International Journal of Advanced Computer Science and Applications, vol. 13, no. 4, pp. 590-598, Apr 2022.
- [17] H. Yamaki, "Efficient cache architecture for table lookups in an internet router," International Journal of Advanced Computer Science and Applications, vol.11, no. 5, pp. 664-672, May 2020.
- [18] S. Even, A. Itai, and A. Shamir, "On the complexity of time table and multi-commodity flow problems," IEEE FOCS, 1975.
- [19] S. Nedeveschi, L. Popa, G. Iannaccone, S. Ratnasamy, and D. Wetherall, "Reducing network energy consumption via sleeping and rate-adaptation," USENIX NSDI, 2008.
- [20] <https://ampl.com/products/solvers/solvers-we-sell/cplex/options/>.
- [21] <https://ampl.com/>.
- [22] M. Gupta, and S. Singh, "Energy conservation with low power modes in Ethernet LAN environments," IEEE INFOCOM, 2007.

Parallel Improved Genetic Algorithm for the Quadratic Assignment Problem

Huda Alfai¹

College of Computer & Information Sciences
Al-Imam Mohammad Ibn Saud Islamic University
Dept. of Computer Science, Riyadh, Saudi Arabia

Yassine Daadaa²

College of Computer & Information Sciences
Al-Imam Mohammad Ibn Saud Islamic University (IMSIU)
Riyadh, Saudi Arabia

Abstract—Quadratic Assignment Problem is one of the most common combinatorial optimization problems that represents many real-life problems. Many techniques are applied to solve Quadratic Assignment Problem, these include exact, heuristic, and metaheuristic methods. A Genetic Algorithm is a powerful heuristic approach used to find optimal solutions or near-to-optimal for Quadratic Assignment problems. In this paper, we developed a Genetic Algorithm with a new crossover operator with new technology closer to that found in nature without a crossover point and a new suggested intelligent mutation operator, then we developed a Parallel Genetic Algorithm using the same crossover and mutation. The sequential Genetic Algorithm will be implemented in the Central Processing Unit (CPU), and the Parallel Genetic Algorithm will be implemented in the Graphical Processing Unit (GPU). This paper presents two comparisons, first calculates elapsed time for crossover, mutation, and selection in both CPU and GPU, then compares the results. This comparison clearly shows the enhancement degree of computation time in the parallel environment, which is around half the time executed in the sequential environment. The second comparison, iterates these operators into several generations, using twenty benchmark instances reported in Quadratic Assignment Problem Library with sizes from (12-70), population size equal to 600, the number of generations equal to 2000, and the maximum number of parallel threads will be 600. Proposed crossover and mutation give the optimal solutions with ten benchmarks with problem sizes from 12 to 32 in both Sequential Genetic Algorithm and Parallel Genetic Algorithm, the next ten benchmarks give solutions closed to the optimal solution with a small error rate.

Keywords—Component; Quadratic Assignment Problem (QAP); Genetic Algorithm (GA); Parallel Genetic Algorithm (PGA); Sequential Genetic Algorithm (SGA); Central Processing Unit (CPU); Compute Unified Device Architecture (CUDA); Quadratic Assignment Problem Library (QAPLIB); Best Known Solution (BKS); Average Percent Deviation (APD)

I. INTRODUCTION

The Quadratic Assignment Problem (QAP) is one of the most common combinatorial optimization problems that represents many real-life problems. The QAP involves the assignment of n facilities that have flows (weights) among them to n possible locations that also have distances among them to achieve the minimum sum of the distances multiplied by flows, this minimum sum will be reached by assigning high facilities to nearby locations and small facilities to far locations. The problem was first introduced as a mathematical

model for economic activities in 1957[1], then it was becoming a fundamental and important problem to represent several applications in different areas, such as computer backboard wiring, locating clinics with a hospital, locating machine and electronic components, assignment of buildings in a university campus, etc.

The quadratic assignment problem (QAP) consists of n facilities and n possible locations, exactly one facility for each location. For each pair of facilities, a flow matrix, $F = [f_{ij}]$ is defined, which consists of flow values that must be required to move from facility i to facility j . Also, for each pair of locations, a distance matrix, $D = [d_{kl}]$ is defined and it consists of distance values between location k to location l . The assignment of facility i is not independent of other assignments, so when assigning facility i to location k we must consider the assignment for all other facilities that have nonzero relationships with facility i . Let $a = \{a(1), a(2), \dots, a(n)\}$ be an assignment, where $a(i)$ represents the location of the facility i . The problem is to assign to each location exactly one facility to minimize the cost of the objective function as shown in Equation. 1.

$$Z_a = \sum_{i=1}^n \sum_{j=1}^n f_{ij} d_{a(i)a(j)} \quad (1)$$

Since the solution is derived from $n!$ possible assignments, it makes the problem impossible to solve in polynomial time with moderate problem size, even with modern computers.

The QAP solving methods can be categorized into three main classifications: exact methods, heuristic methods, and meta-heuristic methods. The exact methods give the exact optimal solution, but the drawback of such methods is the long computational time that makes the solution impossible. Therefore, the problem was restored to be solved using heuristics and meta-heuristic methods which overcome the problem of long computational time, but they also have their drawback. Heuristics and meta-heuristic methods do not guarantee to provide the exact optimal solution, but they instead provide a good solution, near to optimal solution, in reasonable computational time. Genetic algorithms, simulated annealing, tabu search, artificial neural network, etc., are some well-known heuristic methods, and genetic algorithm is considered as one of the best heuristic methods.

A Genetic Algorithm (GA) provides individual candidate solutions that do not hold any dependencies between them, so, it will be easy to implement such an algorithm in parallel to get a more considerable speedup.

This paper uses a parallelism concept which in turn becomes an effective way to simplify the difficult problems and reduce its computational time. Additionally, GA is a popular effective heuristic approach in both computation time and solution quality. So, they have motivated us to take the advantage of both GA and parallelism to solve that difficult problem.

This work exploits the recent improvement in the graphical processing unit (GPU) which is expanded to include parallel computation rather than just graphical purpose. So, we will propose a solution for QAP using a proposed genetic algorithm with enhancement in crossover and mutation. These enhancements are suggested new crossover operator with new technology which closer to that found in nature without crossover point and new intelligent mutation operator which in turn improve solution quality.

II. BACKGROUND AND RELATED WORK

Genetic algorithms were first invented on QAP by John Holland at the University of Michigan in 1975[2]. The first applied for GA in QAP was in 1994 by Fleurent and Ferland[3]. GA is considered as a type of stochastic and local search technique, which are based on three natural operators: selection, crossover, and mutation. Also, there are many recent efficient algorithms, we will present a brief study about them to explore the new techniques and take advantage of them.

Radomil Matousek et al [4] presented Metaheuristic Optimization Using HC12 Algorithm. It is categorized as a parallel algorithm implemented on GPU. It used HC12 which is a Genetic Algorithm using binary encoding which depends on the next population is a population from the current solution neighborhood. This algorithm gives the optimal solutions for 8 problems with sizes (12 -32) in a short run time of an average of 1.89 seconds.

Takeshi Okano et al [5] proposed variant k-opt local search (vKLS) which is categorized as a sequential algorithm in a CPU environment, vKLS used a variable depth approach that depends on exchanging multiple nodes at a time rather than just two nodes. They combine two strategies best-improvement move and the first-improvement move. vKLS tested on 48 QAPLIB instances with a range of 20 - 150 in a fixed period equal to 60 seconds.

Ensieh et al improved the performance of the (NIFLS) Fast Local Search algorithm in the sequential environment by adding Temperature characteristics from simulated annealing to conduct the search to explore the search space wider[6]. The algorithm gets 0.26 APD in average execution time 1207 seconds.

Erdener et al developed ILS (Iterated Local Search) algorithm using GPU parallelism[7]. They implement the multi-start technique, use the delta function instead of

calculating object function for each neighbor and design a mutation operator to escape the local optimum. The algorithm works 6.31 to 11.93 times faster than sequentially one.

Omar Abdelkaf et al. [8] suggested Parallel iterative Tabu Search (PITS) by parallelizing an existing TS algorithm called Ro-Ts using a grid of 5000 CPUs. PITS works with 350 iterations inside the process, 100 global iterations, and 40 processes. PITS gives an average standard deviation equal to 12.19 in average time equal to 13.01 minutes with problems with size 343.

Also Emrullah et al presented an algorithm called the Parallel Simulated Annealing method with multi-start technique (PMSA) using GPU parallelism[8]. PMSA starts the next SA algorithm with the best previous generated value rather than a random permutation, this technique is called the multi-start approach. It provides the optimal solution for 196 instances except for 14 instances in time less than 60 seconds.

Lopez et al presented GA-CPLS algorithm which is a type of CPU level parallelism[9]. CPLS operation depends on a group of nodes called explorers. GA-CPLS performed the Genetic algorithm as the main explorer to generate the population as a head node, other explorer nodes execute the Extremal Optimization Algorithm and robust Tabu search. GA-CPLS gives 0.054 APD on an average time of 82.7 minutes.

Seyda et al improved sequential Hybrid GA called IHGA[10]. Its idea takes from combining genetic algorithm, simulated annealing algorithm, and the greedy algorithm. It enhances the solution by 13.33, 7.94, 2.50, and 0.29 percent better than the greedy algorithm, DA, classical GA, and SA respectively.

Soukaina et al developed a Hybrid Chicken Swarm Optimization (HCSO)[11]. HCSO applies GPU level parallelism and integrates Chicken Swarm Optimization CSO with Greedy Randomized Adaptive Search Procedure GRASP. GRASP run with a 2-opt Local Search for constructing the initial population. HCSO finds the optimal solution for 85% of 30 QAP instances.

Mohamed et al enhanced Whales Optimization Algorithm by integrating it with Tabu Search (WAITS)[12]. WAITS was applied in a sequential environment, and it enhances the speed of convergence and local search inside the Whales Algorithm (WA). WAITS provides the optimal solutions for 86 instances out of 122 instances.

Previous studies explored many recent heuristics and metaheuristics algorithms in solving QAP either in parallel or in a sequential environment. Parallelism can be designed at the CPU level or GPU level. As we see from reviewed algorithms, parallel algorithms designed by GPU produced better results in computational time and algorithms like GA will provide a high-quality solution in a reasonable time. This will motivate us to design a new GA with a new crossover operator with new technology closer to that found in nature, it depends on arranging genes in a specific way without the need for a crossover point, and also suggested an intelligent mutation operator in the GPU environment.

The proposed method will be implemented and tested in a sequential environment and then in parallel to compare results and to show the degree of parallel improvement using benchmark instances available in QAPLIB[13].

This paper was organized into sections, each section treats a part of our works. The second section shows the methodology of our works, the next section illustrates the overall structure of the proposed algorithm, the fourth section analyzes and explores the results, and finally the conclusion.

III. METHODOLOGY

A. Population Initialization Method

Population sets will be initialized randomly concerning the problem size. Additionally, make sure this population does not have incomplete or invalid individuals and all nodes are existing and forming a complete solution. Also, be sure the individual does not have redundant nodes or invalid nodes.

B. Selection Method

The proposed GA applied the selection to two places in the algorithm. First, parents' selection is called the stochastic remainder selection method. It works by assigning a probability to every individual to be chosen as a parent. This method takes each individual's fitness then divides it by average fitness, the integer part of the division represents the number of appearances of the individual as a parent, and the remaining fractional part is used to stochastically fill the remaining parents to stochastic places.

The second application of selection was after crossover operation when deciding about if a current parent will stay for the next generation or be replaced by its best offspring. This type of survivor selection is called the steady-state approach.

C. Crossover Operator

In this paper, we propose a new crossover method that produces an individual who inherits from parent's characteristics as much as possible. This method will preserve the order of the inherited nodes from both parents without making a crossover point.

The following example will illustrate the proposed crossover method by using the facility matrix and distance matrix that is used in the "Hud12" benchmark. If we have two parents parent1 with cost = 1956 and parent2 with cost = 1936 each with size 12, as shown in Fig. 1 and Fig. 2, and offspring will be as shown in Fig. 3.

Parent1:

5	4	12	6	10	9	7	1	8	3	11	2
---	---	----	---	----	---	---	---	---	---	----	---

Fig. 1. Crossover _ parent1

Parent2:

12	6	9	2	4	11	10	1	5	8	7	3
----	---	---	---	---	----	----	---	---	---	---	---

Fig. 2. Crossover _ parent2.

Offspring:

5	4	12	6	10	9	2	11	1	8	7	3
---	---	----	---	----	---	---	----	---	---	---	---

Fig. 3. Crossover_offspring.

There are two indexes (index1= 0) which point to the first index in parent1, (index2=size-1=11) which point to the last index in parent2. Start filling offspring by these two indexes, at the same time, as shown in Fig. 4.

Step1: index1=0, index 2=11, offspring will be:

5											3
---	--	--	--	--	--	--	--	--	--	--	---

Fig. 4. Crossover First Step.

5 is the first node in parent1, 3 is the last node in parent2, increment index 1, decrement index2, index1=1, index2=10.

Step2: index1=1, index2=10, before inserting must check if the new node exists in new offspring if not just insert it, if exist go to the next node in the corresponding parent, offspring will be , as shown in Fig. 5.

5	4								7	3
---	---	--	--	--	--	--	--	--	---	---

Fig. 5. Crossover Second Step.

4 is the second node in parent1, 7 is the second node from the last in parent2, increment index 1, decrement index2, index1=2, index2=9.

Step3: index1=2, index 2= 9, before inserting must check if the new node exists in new offspring if not just insert it, if exist go to the next node in the corresponding parent, offspring will be as shown in Fig. 6:

5	4	12						8	7	3
---	---	----	--	--	--	--	--	---	---	---

Fig. 6. Crossover Third Step.

12 is the third node in parent1, 8 is the third node from the last in parent2, increment index1, decrement index2, index1=3, index2=8.

Step4: index1=3, index 2= 8, before inserting must check if a new node exists in new offspring if not just insert it, if exist go to the next node in the corresponding parent, offspring will be , as shown in Fig. 7.

5	4	12	6					1	8	7	3
---	---	----	---	--	--	--	--	---	---	---	---

Fig. 7. Crossover Fourth Step.

6 is the fourth node in parent1, 5 is the fourth node from the last in parent2 but 5 exists in offspring, so go to the fifth node from the last in parent2 which is 1 then check if doesn't exist in offspring insert 1, increment index1, decrement index2, index1=4, index2=7.

Step5: index1=4, index 2= 7, before inserting must check if the new node exists in new offspring if not just insert it, if exists go to the next node in the corresponding parent, offspring will be as shown in Fig. 8.

5	4	12	6	10			11	1	8	7	3
---	---	----	---	----	--	--	----	---	---	---	---

Fig. 8. Crossover Fifth Step.

Ten (10) is the fifth node in parent1, 10 is the sixth node from the last in parent2 but 10 exists in offspring, so go to the

seventh node from the last in parent2 which is 11 then check if does not exist in offspring insert 11, increment index 1, decrement index2, index1=5, index2=6.

Step 6: index1=5, index 2= 6, before inserting must check if the new node exists in new offspring if not just insert it, if exist go to the next node in the corresponding parent, offspring will appear as shown in Fig. 9.

5	4	12	6	10	9	2	11	1	8	7	3
---	---	----	---	----	---	---	----	---	---	---	---

Fig. 9. Crossover Sixth Step.

Nine (9) is the sixth node in parent1, 4 is the eighth node from the last in parent2 but 14 exists in offspring, so go to the ninth node from the last in parent2 which is 2 then check if doesn't exist in offspring insert 2. The cost for the generated offspring = 1868 which is better than the cost of parents.

Crossover must be simple as possible to achieve maximum utilization of GPU benefits. The generated offspring was produced by simple crossover but inherit many features from parents selected by a strong selection method.

D. Mutation Operators

The proposed GA uses a new mutation operator that works as scanning the individual to find the maximum product (flow * distance) located between facility(i) to the facility (i+1). Then swap facility (i+1) with random node from the individual.

This proposed mutation can be illustrated as shown in the following example: The following individual belongs to the "Had12" benchmark with cost = 1902, as shown in Fig. 10.

4	6	3	1	8	2	9	10	7	12	5	11
---	---	---	---	---	---	---	----	---	----	---	----

Fig. 10. Individual before Mutation.

After applying the mutation operator, the cost will be = 1834, and the individual will be as shown in Fig. 11.

4	6	3	1	7	2	9	10	8	12	5	11
---	---	---	---	---	---	---	----	---	----	---	----

Fig. 11. Individual after Mutation.

IV. STRUCTURE OF THE PROPOSED PARALLEL GENETIC ALGORITHM

The following algorithm shows the general structure of the proposed PGA, followed by a system diagram to represent the PGA structure, as shown in Fig. 12.

PGA exploited graphical processing unit (GPU) for non-graphical parallel computation, the proposed algorithm uses a large single population of individuals which is distributed among several threads in GPU. Each thread performs three GA operators Crossover, Mutation Survivor, and Selection because they are suitable to implement in the parallel environment as shown in Fig. 12. This means, does not need to force threads to communicate between each other or lock other threads, or wait for other threads until unlocking, this parallelism technique maintains data integrity and consistency also threads' waiting time is almost non-existent. The proposed algorithm was shown in Table I.

TABLE I. ALGORITHM I

Algorithm1: proposed Parallel Genetic Algorithm

```

Population Size= N
Problem Size = n
Number of threads =N
Termination condition=2000 generation

Create random Population with size N

while termination condition is not reached do

    calculate fitness values

    Reorder individuals according to stochastic probability

    For each thread i, in parallel do

        Select parent 1 =individual i

        Select parent 2 = individual i+1,

        For each facility j in offspring

            offspring = Assign facility j from left of individual i

            offspring =Assign facility n - j+1 right of individual i+1

        Find maximum product (flow * distance) between facility(k) to
        facility (k+1) in individual i

            offspring = swap facility (k+1) with random picked facility.

        If offspring cost < parent 1 cost

            Replace (individual i = offspring)

    end while
    
```

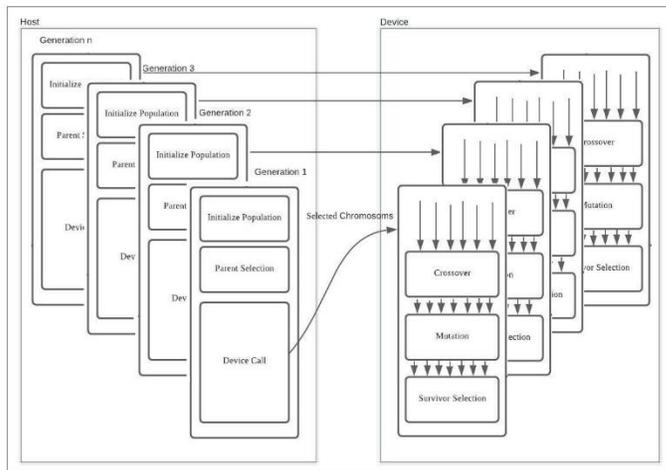


Fig. 12. Overall Proposed PGA Structure.

V. RESULT AND DISCUSSION

This section will show an analysis, discussion, and illustration of the output and results of the proposed method in this paper, results are going to be analyzed in two ways. The

first analysis will present six tables that show elapsed time for GPU and CPU during the execution of proposed crossover, mutation, and selection. The second analysis presents a test of the proposed method in GPU and CPU after embedding it inside several iterations (generations).

The proposed method was tested in CPU of type intel® core™ i7-8565U CPU @ 1.80GHz (8 CPUs) and GPU of kind NVIDIA GeForce MX250 using both CUDA (Compute Unified Device Architecture) and C++ programming languages.

The following four figures show a comparison between CPU and GPU using common QAP benchmarks, while N means the size of population, CPU and GPU time is measured in milliseconds.

A. First Test Illustration

This test shows elapsed time for GPU and CPU during the execution of proposed crossover, mutation, and selection.

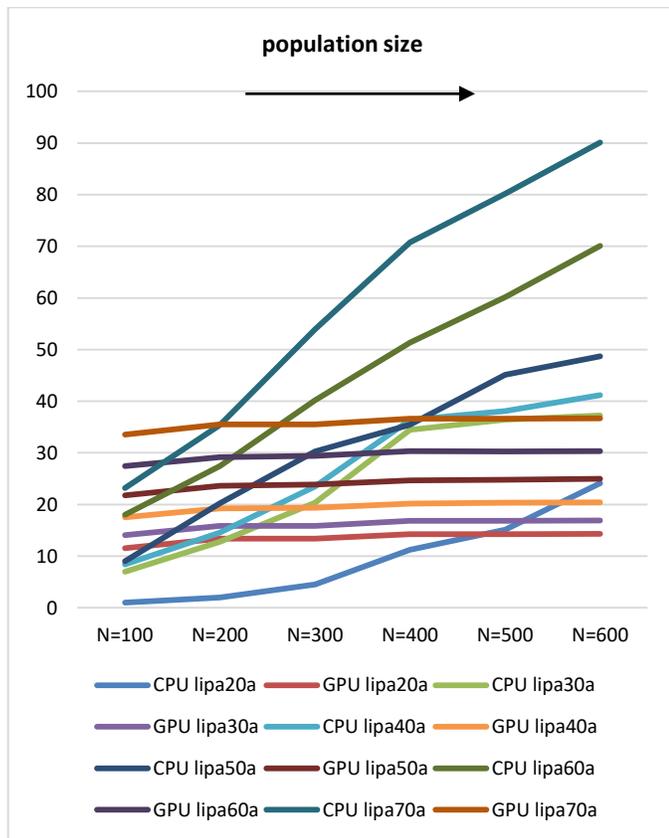


Fig. 13. CPU and GPU Time Illustration with Problem Sizes.

1) For the “lipa20a” benchmark: “lipa20a” benchmark with problem size equal to 20. We notice that when population size equal to 100, 200, 300, and 400 CPU show better results than GPU. Here the problem size is small, and we will only see the enhancement in GPU when the size of the problem and population increase. After increasing the population size to 600 we will observe the GPU enhancement and CPU time become approximately twice the time of GPU. Fig. 13 shows a graphical representation of this problem.

2) For the “lipa30a” benchmark: The enhancement on GPU begins at N=300, then the CPU time will take an increasing rate when population size increases. Compared to GPU time, GPU time does not take a significantly increasing rate while the population size increases, it just took a small increasing rate ≈ of 0.56 milliseconds, as shown in Figure IV.1. CPU continues increasing until it reaches more than 2x time of GPU time at N= 600. Fig. 13 Shows a graphical representation of this problem.

3) For the “lipa40a” benchmark: The improvement on GPU starts when N=300, then the CPU time will increase when population size increases until it reaches nearly 2x the time of GPU at N= 600. On the other hands, GPU time takes a small increasing rate ≈ of 0.58 while the population size increases. Fig. 13 shows a graphical representation of this problem.

4) For the “lipa50a” benchmark: shows the same result as “lipa40”. The CPU looks better than GPU when N=300, but after that, it becomes worse when N>300. CPU becomes around 2x time of GPU at N= 600. As we noted earlier GPU time is not affected much by population increase, as shown in Fig. 13.

5) For “lipa60a” and “lipa70a” benchmarks: CPU looks worse than GPU when population size > 200, it becomes around 2.3x time of GPU at problem size =60 and N= 600 and it becomes around 2.5x time of GPU at problem size = 70 and population size N= 600, as shown in Fig. 13.

6) Fig. 14: shows the worst CPU time state when population size = 600, it becomes around 2x GPU time and it increases at a significant rate when the problem size increases.

B. Second Test

Table II presents a test set using a group of common QAP benchmarks in size range (12 - 70) and population size N =600. After inserting the proposed tested method into the whole Genetic Algorithm program, we will show their elapsed CPU and GPU time after 2000 generations then show the best solution found in both CPU and GPU. Also, measure Average Percent Deviation to measure how much the solution is closed to best known solution (BKS), as shown in Equation. 2.

$$APD = \frac{\text{solution} - \text{BKS}}{\text{BKS}} \quad (2)$$

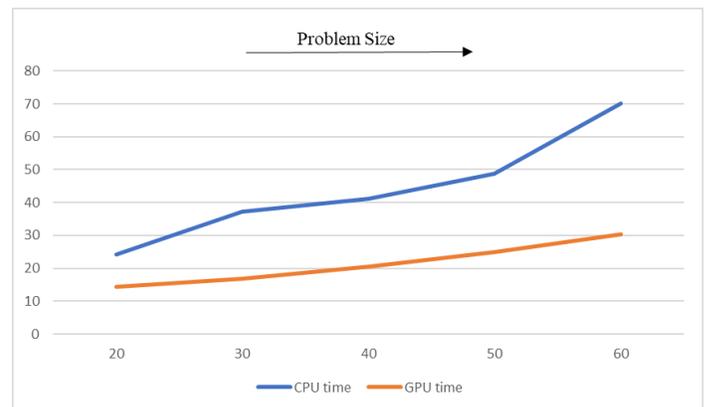


Fig. 14. CPU and GPU Time when Increased Problem Size and N=600.

TABLE II. SOLUTIONS AND ELAPSED TIME FOR CPU AND GPU

Problem Instances	Avg. Whole CPU program time	Avg. Whole GPU program time	CPU solution	GPU solution	Optimal solution	CPU APD	GPU APD
Esc16a	13112.34	22242.32	68	68	68	0	0
Esc16b	7565.65	22133.45	292	292	292	0	0
Esc16c	8301.25	21650.29	160	160	160	0	0
Esc16d	8986.33	21722.73	16	16	16	0	0
Esc16e	8227.89	22620.11	28	28	28	0	0
Esc16g	17005.75	21929.62	26	26	26	0	0
Esc32g	29679.09	31093.71	6	6	6	0	0
Had12	4577.27	19788.06	1652	1652	1652	0	0
Had14	5973.67	20853.74	2724	2724	2724	0	0
Had16	14555.79	21911.14	3720	3726	3720	0	0
Lipa20a	11250.61	24875.72	3797	3809	3683	0.03	0.03
Lipa30a	51688.26	34824.34	13555	13565	13178	0.03	0.03
Lipa40a	60049.76	48419.51	32340	32346	31538	0.03	0.03
Lipa50a	70703.40	65243.23	63476	63557	62093	0.02	0.02
Lipa60a	142518.61	85710.54	109388	109429	107218	0.02	0.02
Lipa70a	133348.29	110122.11	172941	172894	169755	0.02	0.02
Sko42	95743.75	51448.87	16900	17286	15812	0.07	0.09
Sko49	124436.57	62790.45	26719	27317	23386	0.14	0.17
Sko56	176624.76	77348.33	37858	38412	34458	0.1	0.11
Sko64	215862.65	98184.77	53072	53504	45736	0.16	0.17

The proposed crossover and mutation give the optimal solution with the first ten benchmarks with problem sizes from 12 to 32 in both sequential Genetic Algorithm and Parallel Genetic Algorithm, next ten benchmarks give a solution close to the optimal solution with a small error rate, CPU and GPU time are measured in milliseconds.

VI. CONCLUSION

After this study, we found that the GPU time is not affected much by increasing either population size or problem size compared to CPU time. GPU time was increased by a small rate \approx of 0.61 when increasing population size and take a small rate \approx of 4.5 when the size of the problem increases. Also, the CPU time shows its worst when the population size = 600, it becomes around 2x GPU time, and it increases at a significant rate when the problem size increases.

This paper concentrates on applying proposed GA to QAP, which in turn, gives a successful result in finding optimal solutions or solutions near to optimal.

Also, this paper applies proposed GA in the parallel environment which shows a good result in execution time enhancement. As mentioned before, the proposed solution uses a large population size; therefore, a lot of synchronous threads will be needed. So, as future works, we can increase the number of threads by increasing the size of the screen card (NVIDIA). Furthermore, the proposed PGA can be generalized

to cover other optimization problems such as TSP (traveling salesman problem) or VRP (Vehicle routing problem).

As a future work we try to enhance some drawbacks that make algorithm slower such as sequential parent selection, it needs to convert to be work in parallel environment.

REFERENCES

- [1] T. Koopmans and M. Beckmann, "Assignment problems and the location of economic activities," *Econometrica: Journal of the Econometric Society*, vol. 25, no. 1. pp. 53–76, 1957, doi: 10.2307/1907742.
- [2] M. Melanie, "An Introduction to Genetic Algorithms," Cambridge, Massachusetts London, England, Fifth printing, 3, pp. 62–75, 1999.
- [3] C. Fleurent and J. Ferland, "Genetic hybrids for the quadratic assignment problem," *Anon. DIMACS Ser. Discret. Math. Theor. Comput. Sci.*, vol. 16, pp. 173–187, 1994.
- [4] R. Matousek, L. Dobrovsky, and J. Kudela, "The quadratic assignment problem: Metaheuristic optimization using HC12 algorithm," *GECCO 2019 Companion - Proc. 2019 Genet. Evol. Comput. Conf. Companion*, pp. 153–154, 2019, doi: 10.1145/3319619.3322088.
- [5] T. Okano, K. Katayama, K. Kanahara, and N. Nishihara, "A Local Search Based on Variant Variable Depth Search for the Quadratic Assignment Problem," 2018 IEEE 7th Glob. Conf. Consum. Electron. GCCE 2018, pp. 390–391, 2018, doi: 10.1109/GCCE.2018.8574497.
- [6] E. Mohassesian and B. Karasfi, "A new method for improving the performance of fast local search in solving QAP for optimal exploration of state space," 7th Conf. Artif. Intell. Robot. IRANOPEN 2017, pp. 64–72, 2017, doi: 10.1109/RIOS.2017.7956445.
- [7] E. Ozcetin and G. Ozturk, "A Parallel Iterated Local Search Algorithm on GPUs for Quadratic Assignment Problem," vol. 4, no. 2, pp. 123–128, 2018.

- [8] O. Abdelkafi, B. Derbel, and A. Liefoghe, "A Parallel Tabu Search for the Large-scale Quadratic Assignment Problem," 2019 IEEE Congr. Evol. Comput. CEC 2019 - Proc., pp. 3070–3077, 2019, doi: 10.1109/CEC.2019.8790152.
- [9] J. Lopez, D. Munera, D. Diaz, and S. Abreu, "On integrating population-based metaheuristics with cooperative parallelism," Proc. - 2018 IEEE 32nd Int. Parallel Distrib. Process. Symp. Work. IPDPSW 2018, pp. 601–608, 2018, doi: 10.1109/IPDPSW.2018.00100.
- [10] S. M. Turkkahraman and D. Oz, "An Improved Hybrid Genetic Algorithm for the Quadratic Assignment Problem," pp. 86–91, 2021, doi: 10.1109/ubmk52708.2021.9558978.
- [11] S. C. Bourki Semlali, M. Essaid Riffi, and F. Chebihi, "Hybrid chicken swarm optimization with a GRASP constructive procedure using multi-Threads to solve the quadratic assignment problem," Int. Conf. Multimed. Comput. Syst. -Proceedings, vol. 2018-May, 2018, doi: 10.1109/ICMCS.2018.8525992.
- [12] M. Abdel-Basset, G. Manogaran, D. El-Shahat, and S. Mirjalili, "Integrating the whale algorithm with Tabu search for quadratic assignment problem: A new approach for locating hospital departments," Appl. Soft Comput. J., vol. 73, pp. 530–546, 2018, doi: 10.1016/j.asoc.2018.08.047.
- [13] R. E. Burkard, E. Çela, S. E. Karisch, and F. Rendl, "QAPLIB - A Quadratic Assignment Problem Library," J. Glob. Optim., vol. 10, no. 4, pp. 391–403, 1997.

Modeling for Car Quality Complaint Classification based on Machine Learning

Chen Xiao Yu¹, Hou Xia²

Computer School
Beijing Information Science &
Technology University
Beijing, China

Zhang Xiao Min³

Academy of Agricultural Planning
and Engineering
Ministry of Agriculture and Rural
Affairs, Beijing, China

Song Ying⁴

Computer School
Beijing Information Science &
Technology University
Beijing, China

Abstract—Cars play an important role in many aspects of people's social life, and the effective handling of car quality complaints is of great significance to the proper running of cars and the reputation maintenance of car brands; effective classification of car quality complaint texts is the basis of the efficient handling of corresponding quality complaints, while relying on manual classification has disadvantages such as heavy workload, experience dependence, and error proneness; machine learning methods have been quite widely used in the automatic classification modeling for different types of natural language texts. It is of great practical significance to construct the automatic classification model of car quality complaints based on machine learning. Based on the characteristics of car quality complaint texts, this study vectorized the texts after word segmentation, performed feature selection and dimension reduction based on correlation analysis, and combined the progressive model training method and support vector machine to construct the classification model; in model reliability analysis, it was evaluated based on the effect of data amount on the modeling and the effect of text length on the prediction probability distribution. The results show that based on the method in this study, effective automatic classification model of car quality complaint texts could be constructed.

Keywords—Car; quality complaint; natural language text; classification modeling; machine learning

I. INTRODUCTION

The studies on text classification are quite extensive, but there are few related studies on complaint text, and the applicability of classification methods is closely related to text characteristics. The composition and structure of cars are relatively complex; during the long-term use of cars, quality problems might gradually appear, reasonable handling for the quality problems has important effect on the normal operation of cars and the maintenance of user experience, which is also an important decision-making influence factor for people choosing car brand and car product.

Machine learning has been quite extensively applied to natural language text classification in recent years [1-4]. Text classification based on machine learning mainly involves two core links: text vectorization and classification modeling. The methods used in text vectorization mainly include the methods based on word frequency [5-7], the methods based on distributed static word vectors [8-11], and the methods based on distributed dynamic word vectors [12-13]. The methods

used in the classification modeling mainly include classical machine learning methods [14], various neural networks [15-19], ensemble learning [20] and so on.

II. TECHNICAL ROUTE

The technical route of this study includes seven parts, including data sorting, data characteristic analysis, word segmentation, feature extraction, classification modeling, model reliability analysis, summary and prospect.

The part of data sorting includes the acquisition of basic data, and the construction of research dataset based on the text characteristics and research purposes. The part of data characteristic analysis conducts a comprehensive overview of the dataset mainly from the aspects of data distribution, text length characteristics, the distribution of car type, the distribution of purchase time, and the distribution of car brand.

The research object of this study is Chinese text and the study involves word segmentation. The word segmentation part in the technical routes include using Jieba for word segmentation, removal of stop words, word frequency distribution analysis, classification feature word analysis, etc. The removal of stop words aims mainly at removing function words which have little significance for classification, such as the connectives in complaint texts. Word frequency distribution analysis mainly analyze the discrimination and contribution potential of high-frequency words in the classification of car quality complaints, from the perspective of the word frequency distribution of global high-frequency words in different categories. Classification feature word analysis mainly analyzes the characteristics of high-frequency words in each category after removing stop words, and conducts preliminary data status analysis.

The feature extraction part mainly involves three links: text vectorization, feature correlation analysis, and feature selection. In the text vectorization process, the text data is converted to vector form based on bag-of-word method, which doesn't include stop words. In the feature correlation analysis link, the frequency correlation of word features is analyzed through correlation matrix constructing, and the feature selection is performed by removing highly correlated word features to reduce vector dimension so as to improve the efficiency of modeling and classification.

In the classification modeling part, the progressive strategy is used, the proportion of the features used in modeling is gradually increased in multiple stages; the modeling effects under different proportions of features are compared to obtain the optimal modeling feature quantity. The meaning of the progressive strategy is that too few features might don't contain enough necessary information for building an effective classification model, at the same time, too many features might confuse the core information and reduce the classification ability of the model, furthermore, too many modeling features would also result in negative effects on the efficiency of modeling and classification. The classification modeling uses the method of support vector machine, and the evaluation of modeling effect is analyzed from two aspects: the classification quality on the whole dataset and the quality on different categories of complaint texts.

The model reliability analysis part includes three aspects: the effects of data amount on the overall modeling indexes, the effects of data amount on the classification effect in each category, and the effects of text length on the probability distribution of the classification prediction. The amount of training data commonly has an important effect on the reliability of the model, too little data might don't be enough to train a reliable and stable model, and the model's predict ability to new data outside the research dataset might be insufficient or unstable, generally, based on more data, more stable model could be obtained; at the same time, after the amount of data reaches a certain threshold, the continued increase of the data amount commonly no longer has significant effect on the stability of the model. For different types of texts, the amount of data required for classification model training commonly varies. This study analyzes the effect of data amount on the classification modeling of car quality complaint texts by incrementally adding of data and comparing multiple rounds of model training; the evaluation and analysis are carried out from two aspects: the effect of data amount on the overall indexes of modeling, and the effect of data amount on the classification effect in each category. In addition, the text length might have effect on the text classification prediction effect, and the discrimination of the classification prediction probability distribution could reflect to some extent the reliability of the classification prediction, therefore, in this study, the effect of text length on the probability distribution of classification prediction is regarded as another aspect of the reliability evaluation.

At the end of the study, the results are summarized to obtain effective conclusions, and the deficiencies of the study are analyzed, so as to provide reference for subsequent related research and application.

The technical route of this study is shown in Fig. 1.

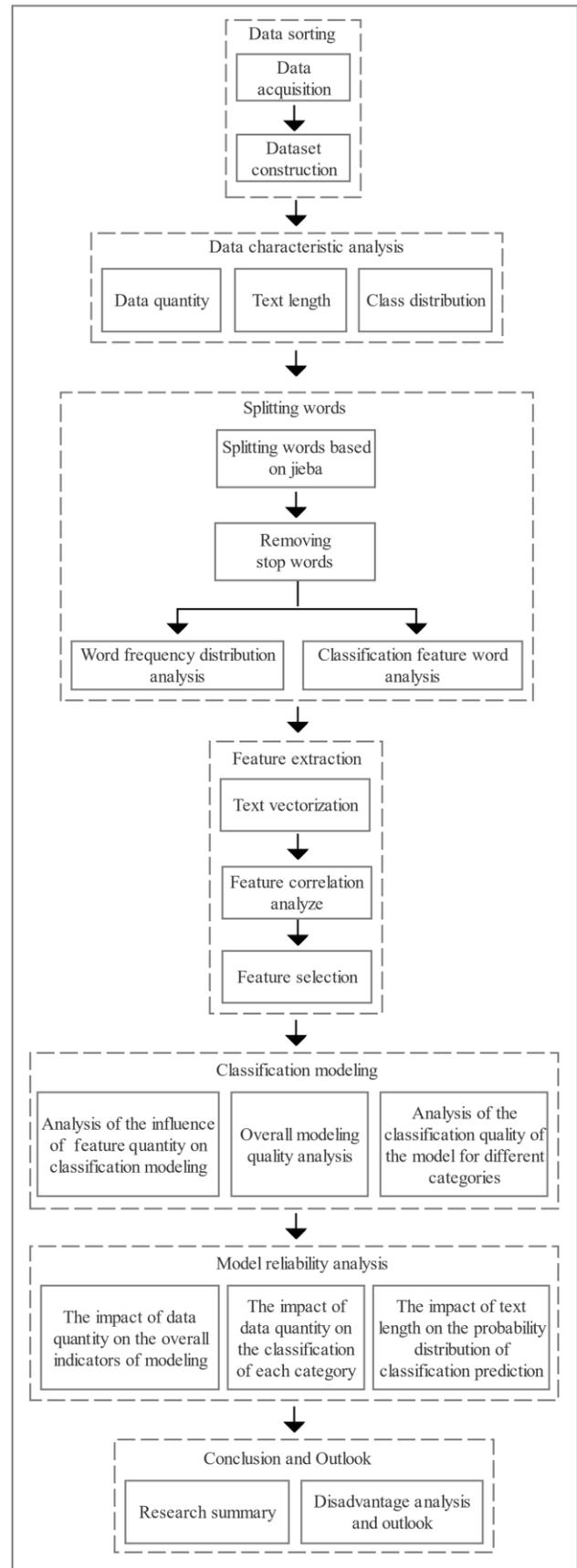


Fig. 1. Technical Route.

III. DATA

The research data of this study comes from the Beijing Car Quality Net Information Technology Limited Company. The dataset of this paper includes 8 categories of car quality complaint text data, including engine/electric motor, transmission, clutch, steering system, braking system, tires, front and rear axles and suspension system, car body accessories and electrical appliances. The data amount is 2400,

and for every category, the data amount is 300. The data amount and text length characteristics are shown in Table I.

The car quality complaint texts involve attribute labels such as car type, purchase time, car brand, etc., and the attribute differences might influence the classification model training and the texts classification prediction. The data distribution of the dataset used in this paper in terms of car type, purchase time, and car brand is shown in Fig. 2.

TABLE I. DATA DESCRIPTION

No.	Category	Data amount	Average length of text	Maximum length of text	Minimum length of text	Text length standard deviation
1	Engine / electric motor	300	18.0267	27	14	1.8194
2	Transmission	300	17.8567	25	14	1.8332
3	Clutch	300	17.9000	24	15	1.7436
4	Steering system	300	17.7767	23	15	1.7214
5	Braking system	300	17.9233	23	14	1.7323
6	Tires	300	17.6500	24	15	1.7216
7	Front and rear axles and suspension system	300	17.9467	23	15	1.7571
8	Car body accessories and electrical appliances	300	18.0767	26	14	2.0845
9	All	2400	17.8946	27	14	1.8071

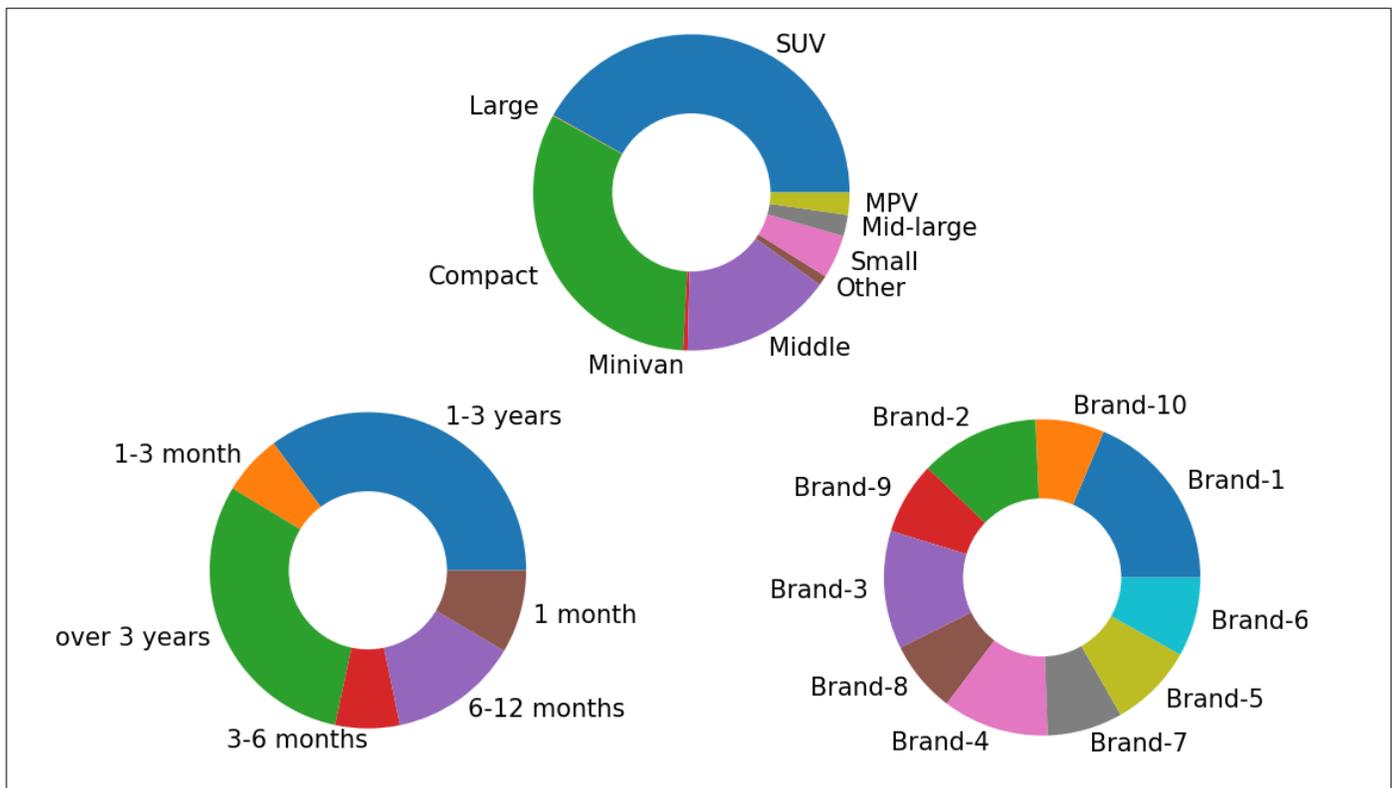


Fig. 2. Data Distribution Characteristics.

IV. SPLITTING WORDS

The research data of this study is Chinese text, and it is necessary to divide texts into words. This study uses the Jieba word segmentation tool which is widely used in the field of Chinese word segmentation to separate the words; the statistics and analysis for word segmentation results are carried out from the aspects of category, number of characters, number of separated words, number of unique words, repetition rate, etc. The word segmentation results are shown in Table II.

Fig. 3 depicts the word frequency distribution of the global high frequency words in different categories. The difference of

the frequency distribution in different categories of the global high frequency words is an important reference factor for the evaluation of the potential classification discrimination contribution ability of these words. If there are widely significant differences in the word frequency distribution of global high-frequency words in different categories, the method based on word frequency might have well applicability for corresponding text classification modeling scene.

After the word segmentation and the removal of stop words, the top 20 high-frequency feature words of each category are shown in Table III.

TABLE II. WORD SEGMENTATION RESULTS

No.	Category	Number of characters	Number of separated words	Number of unique words	Repetition rate
1	Engine / electric motor	5408	2576	584	0.7733
2	Transmission	5357	2457	499	0.7969
3	Clutch	5370	2512	413	0.8356
4	Steering system	5333	2593	479	0.8153
5	Braking system	5377	2610	495	0.8103
6	Tires	5295	2655	285	0.8927
7	Front and rear axles and suspension system	5384	2615	477	0.8176
8	Car body accessories and electrical appliances	5423	2603	766	0.7057
9	All	42947	20621	1866	0.9095

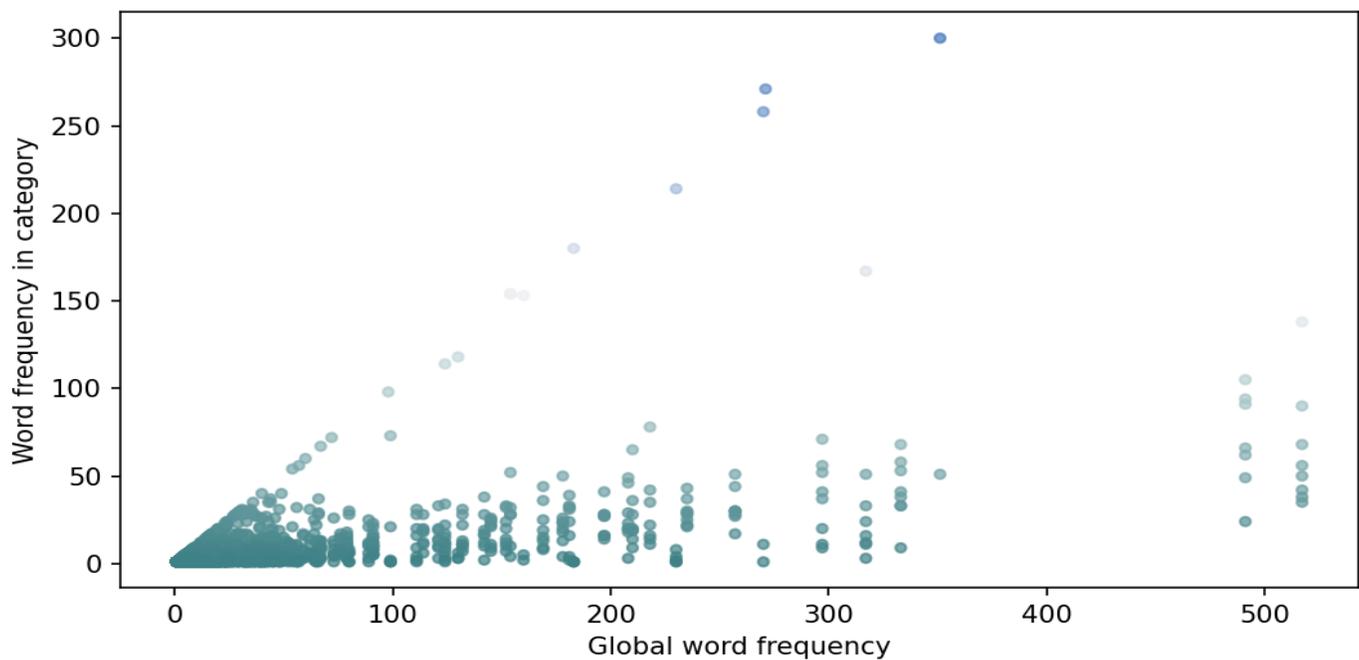


Fig. 3. Word Frequency Distribution.

TABLE III. CATEGORY HIGH FREQUENCY FEATURE WORDS

No.	Engine / electric motor	Transmission	Clutch	Steering system	Braking system	Tires	Front and rear axles and suspension system	Car body accessories and electrical appliances
1	Engine	Gearbox	Clutch	Turn	Brake	Tire	Factory	*
2	Resolve	Fault	Factory	Steering wheel	Factory	*	*	*
3	*	Factory	Hope	Factory	Resolve	*	Eat tires	Rust
4	Engine oil	*	Cause	Resolve	*	*	Driving	Resolve
5	Abnormal noise	Driving	Factory	*	Fault	*	Partial wear	Battery
6	Burn	*	When	Driving	Driving	*	Shock absorber	Power outage
7	*	Resolve	Resolve	Hope	*	Peeling	Oil spill	Hope
8	Fault	When	Driving	When	When	Cracked	*	Factory
9	*	Setback	Shift	Factory	Malfunction	Skin	Tire	*
10	Start up	Shift	*	Stuck	Handbrake	*	*	Start up
11	Hope	Oil spill	Oil spill	*	ABS	Change	Resolve	Body
12	Factory	When	Start	When	Hope	*	Rear wheel	Cause
13	*	Speed up	Jitter	Caton	Electronic	Affect	Change	Change
14	Particles	*	*	Not yet	Jitter	Factory	Chassis	*
15	Blockage	Electromechanical	*	Steering machine	*	Hope	*	Factory
16	*	Unit	*	*	Wear	*	Hope	*
17	*	*	Invalid	Direction	*	Drum kit	*	Fault
18	Oil spill	*	*	*	Factory	*	Cause	*
19	Catch	*	*	Help	*	*	Factory	Cracked
20	Device	Factory	Pedal	Affect	Cause	*	*	*

V. FEATURE EXTRACTION AND CLASSIFICATION MODELING

This study uses bag-of-word method which is based on word frequency for text vectorization; the correlation of features is analyzed based on correlation matrix; and feature selection is performed based on feature correlation to reduce the dimension of text vectors and improve the efficiency of classification model training and text classification. The correlation heatmap of the global high-frequency words after removing stop words is shown in Fig. 4. Due to space limitations, Fig. 4 only shows the relevance of the top 15 high-frequency words in the global word frequency.

This study uses a progressive feature selection strategy to incrementally set the feature usage ratio; the classification model is trained based on the SVM method, and the modeling results are evaluated and analyzed from the perspectives of overall accuracy, overall recall, overall F value, and F value of each category. The progressive feature selection strategy is beneficial to obtain a reasonable threshold of the model feature quantity, if too few features are used, the modeling effect might be adversely affected due to insufficient information, while if too many features are used, the model quality, model training efficiency, and classification prediction efficiency might be adversely affected due to the introduction of non-core information confusion. The training results of the classification model are shown in Table IV.

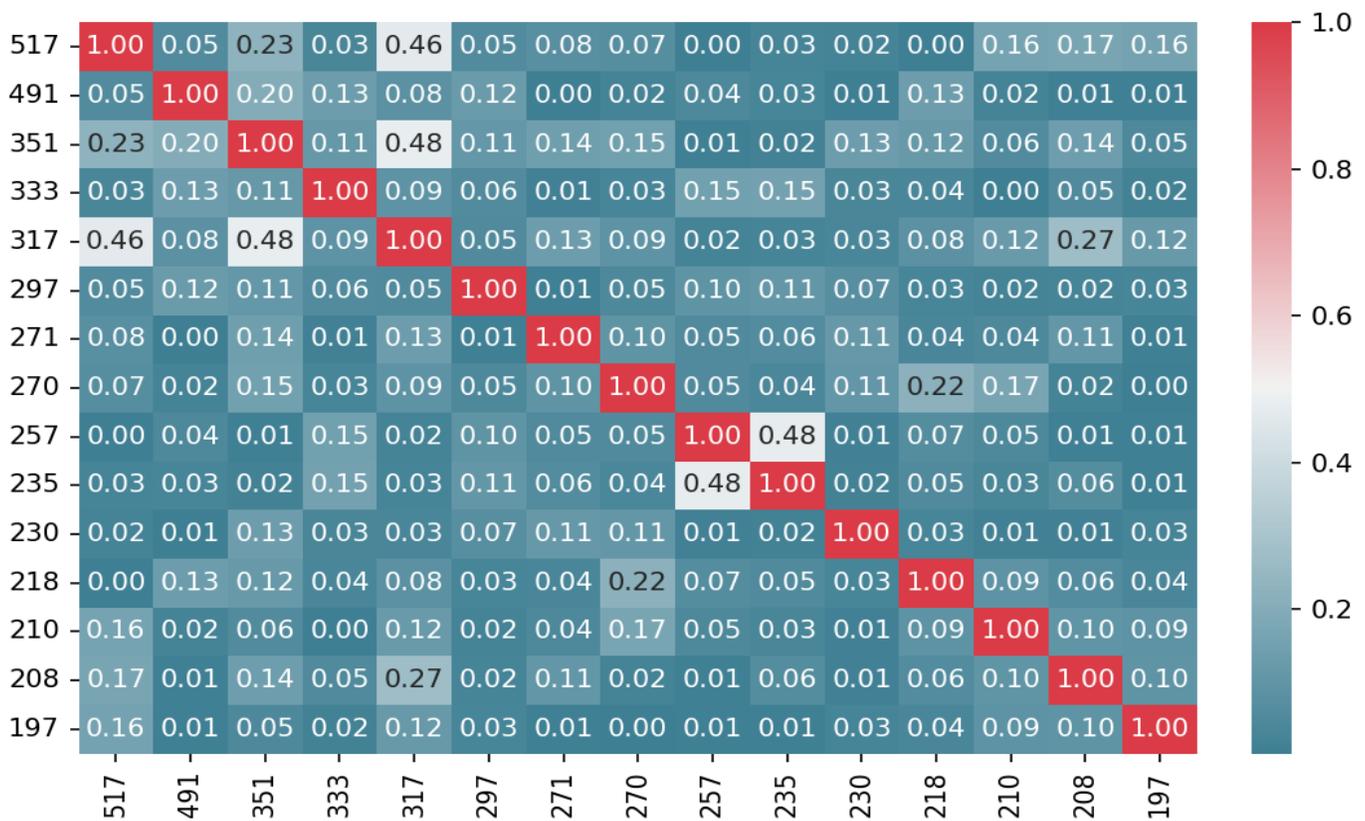


Fig. 4. Correlation Matrix Heatmap.

TABLE IV. CLASSIFICATION MODEL TRAINING

No.	Feature selection ratio	Feature amount	Accuracy	Precision	Recall	F1-score	Highestf1-scoreofeachcategory	Lowestf1-scoreofeachcategory
1	10%	156	0.8958	0.8997	0.8958	0.8968	1.0000	0.7500
2	20%	313	0.8958	0.8997	0.8958	0.8968	1.0000	0.7500
3	30%	469	0.8958	0.8997	0.8958	0.8968	1.0000	0.7500
4	40%	626	0.8958	0.8997	0.8958	0.8968	1.0000	0.7500
5	50%	782	0.8958	0.8997	0.8958	0.8968	1.0000	0.7500
6	55%	860	0.9375	0.9422	0.9375	0.9380	1.0000	0.8788
7	60%	938	0.9375	0.9443	0.9375	0.9384	1.0000	0.8657
8	65%	1017	0.9292	0.9340	0.9292	0.9300	1.0000	0.8485
9	70%	1095	0.9250	0.9296	0.9250	0.9259	1.0000	0.8485
10	75%	1173	0.9250	0.9296	0.9250	0.9259	1.0000	0.8485
11	80%	1251	0.9250	0.9296	0.9250	0.9259	1.0000	0.8485
12	85%	1329	0.9250	0.9296	0.9250	0.9259	1.0000	0.8485
13	90%	1408	0.9167	0.9220	0.9167	0.9178	1.0000	0.8358
14	95%	1486	0.9167	0.9208	0.9167	0.9176	1.0000	0.8475
15	100%	1564	0.9167	0.9238	0.9167	0.9181	1.0000	0.8235

VI. MODEL RELIABILITY ANALYSIS

Model reliability analysis is of great significance to the evaluation of model quality. This study analyzes the reliability of the model from two aspects: the effect of data amount on the classification modeling and the effect of text length on the classification prediction. The training data amount commonly has direct effect on the reliability and stability of text classification model, too little data might lead to limited applicability of the trained model and unstable prediction ability for new data, after the amount of model training data reaches a certain value, the effect of incremental data on the model training effect is commonly no longer significant.

Based on incremental data setting, this study compared multiple rounds of text classification model training, and the result parameters are shown in Table V. Text length is an important factor in text classification model training and classification prediction, the difference of the probability distribution in the classification prediction for different lengths of texts is another effective measure of the reliability of the classification model. In this study, the first 8 texts and the last 8 texts in the global ranking of text length are selected to analyze the probability distribution in the classification prediction; the results are shown in Table VI.

TABLE V. THE EFFECT OF DATA AMOUNT ON MODEL TRAINING

No.	Data using ratio	Data amount	Accuracy	Precision	Recall	F1-score	Highest f1-score of each category	Lowest f1-score of each category
1	10%	240	0.8333	0.8438	0.8333	0.8321	1.0000	0.3333
2	20%	480	0.8333	0.8438	0.8333	0.8321	1.0000	0.3333
3	30%	720	0.8333	0.8438	0.8333	0.8321	1.0000	0.3333
4	40%	960	0.8333	0.8438	0.8333	0.8321	1.0000	0.3333
5	50%	1200	0.8333	0.8438	0.8333	0.8321	1.0000	0.3333
6	55%	1320	0.9015	0.9080	0.9017	0.8994	1.0000	0.7333
7	60%	1440	0.9236	0.9272	0.9236	0.9230	1.0000	0.8333
8	65%	1560	0.9231	0.9304	0.9230	0.9230	1.0000	0.8095
9	70%	1680	0.9405	0.9420	0.9405	0.9399	1.0000	0.8571
10	75%	1800	0.9333	0.9338	0.9331	0.9318	1.0000	0.8372
11	80%	1920	0.9323	0.9330	0.9323	0.9322	1.0000	0.8085
12	85%	2040	0.9265	0.9272	0.9262	0.9255	1.0000	0.8302
13	90%	2160	0.9398	0.9428	0.9398	0.9401	1.0000	0.8627
14	95%	2280	0.9386	0.9445	0.9383	0.9393	0.9825	0.8923
15	100%	2400	0.9375	0.9443	0.9375	0.9384	1.0000	0.8657

TABLE VI. THE PREDICTION PROBABILITY DISTRIBUTION OF THE FIRST 8 AND LAST 8 TEXTS IN THE GLOBAL RANKING OF TEXT LENGTH

No.	Engine / electric motor	Transmission	Clutch	Steering system	Braking system	Tires	Front and rear axles and suspension system	Car body accessories and electrical appliances
F-1	0.9236	0.0212	0.0017	0.0030	0.0034	0.0027	0.0032	0.0411
F-2	0.9934	0.0018	0.0002	0.0002	0.0023	0.0004	0.0004	0.0012
F-3	0.0029	0.0051	0.0015	0.0013	0.0027	0.0019	0.0032	0.9813
F-4	0.0102	0.9530	0.0227	0.0019	0.0007	0.0005	0.0024	0.0086
F-5	0.0433	0.0225	0.0040	0.0063	0.0133	0.0037	0.0259	0.8810
F-6	0.9988	0.0002	0.0003	0.0001	0.0001	0.0003	0.0000	0.0001
F-7	0.0247	0.9491	0.0161	0.0004	0.0005	0.0012	0.0027	0.0052
F-8	0.0066	0.0006	0.9775	0.0017	0.0037	0.0023	0.0027	0.0050
L-1	0.9789	0.0009	0.0049	0.0008	0.0042	0.0007	0.0013	0.0083
L-2	0.9420	0.0015	0.0008	0.0022	0.0070	0.0022	0.0044	0.0400
L-3	0.9352	0.0102	0.0060	0.0015	0.0030	0.0023	0.0092	0.0325
L-4	0.0247	0.9403	0.0108	0.0026	0.0035	0.0024	0.0044	0.0113
L-5	0.0047	0.9879	0.0034	0.0003	0.0005	0.0010	0.0007	0.0015
L-6	0.0015	0.9848	0.0101	0.0003	0.0005	0.0003	0.0010	0.0015
L-7	0.0005	0.0001	0.0003	0.0005	0.9946	0.0008	0.0008	0.0023
L-8	0.0231	0.0044	0.0018	0.0023	0.0058	0.0024	0.0029	0.9575

VII. CONCLUSION AND OUTLOOK

This study focuses on the automatic classification of car quality complaint, the research content mainly includes data characteristic analysis, word segmentation, feature extraction, classification modeling, and model reliability analysis. The research results show that based on the method combining the text vectorization based on word frequency, the feature selection and dimensionality reduction based on correlation analysis, and the feature increase SVM model training, the effective classification model for car quality complaints texts could be obtained; in this study, the best modeling effect is obtained when using 938 features for model training, among the global indexes, the accuracy, recall, and f1-score reach 0.9375, 0.9375, and 0.9384 respectively, the highest f1-score of each category is 1.0000, and the lowest is 0.8657; in the model reliability evaluating based on incremental data amount, after the data using ratio reaches 75%, the training effect is almost stable, the classification prediction probability distribution analysis based on the global long texts and global short texts shows that the classification probability values obtained from the model shows a high degree of discrimination overall.

In general, based on the method of this study, effective modeling for the automatic classification of car quality complaint texts could be realized; at the same time, the research content of this study belongs to theoretical research which has not been applied to practice, it is expected that this study could provide effective reference for subsequent research and practical application.

REFERENCES

- [1] Zhu Fang Peng, Wang Xiao Feng, Text classification for ship industry news [J], Journal of Electronic Measurement and Instrumentation, 2020, 34 (01): 149-155.
- [2] Zhao Ming, Du Hui Fang, Dong Cui Cui, Chen Chang Song, Diet health text classification based on word2vec and LSTM [J], Transactions of the Chinese Society for Agricultural Machinery, 2017, 48 (10): 202-208.
- [3] Bao Xiang, Liu Gui Feng, Yang Guo Li, Patent text classification method based on multi-instance Learning [J], Information Studies: Theory & Application, 2018, 41 (11): 144-148.
- [4] Wen Chao Dong, Zeng Cheng, Ren Jun Wei, Zhang Yan, Patent text classification based on ALBERT and bidirectional gated recurrent unit [J], Journal of Computer Applications, 2021, 41 (02): 407-412.
- [5] Hu Jing, Liu Wei, Ma Kai, Text categorization of hypertension medical records based on machine learning [J]. Science Technology and Engineering, 2019, 19 (33): 296-301.
- [6] Yu Hang, Li Hong Lian, Lü Xue Qiang, Text classification of NPC report contents [J], Computer Engineering and Design, 2021, 42 (06): 1772-1778.
- [7] Wang Xiang Xiang, Fang Hui, Chen Chong Cheng, Classification technique of cultural tourism text based on naive Bayes [J]. Journal of Fuzhou University (Natural Science Edition), 2018, 46 (05): 644-649.
- [8] Zhou Qing Hua, Li Xiao Li, Research on short text classification method of railway signal equipment fault based on MCNN [J]. Journal of Railway Science and Engineering, 2019, 16 (11): 2859-2865.
- [9] Feng Shuai, Xu Tong Yu, Zhou Yun Cheng, Zhao Dong Xue, Jin Ning, et al. Rice knowledge text classification based on deep convolution neural network [J]. Transactions of the Chinese Society for Agricultural Machinery, 2021, 52 (03): 257-264.
- [10] Niu Zhen Dong, Shi Peng Fei, Zhu Yi Fan, Zhang Si Fan, Research on classification of commodity ultra-short text based on deep random forest [J]. Transactions of Beijing Institute of Technology, 2021, 41 (12): 1277-1285.
- [11] Zhang Yu, Liu Kai Feng, Zhang Quan Xin, Wang Yan Ge, Gao Kai Long, A combined-convolutional neural network for Chinese news text classification [J]. Acta Electronica Sinica, 2021, 49 (06): 1059-1067.
- [12] Li Ke Yue, Chen Yi, Niu Shao Zhang, Social E-commerce text classification algorithm based on BERT [J], Computer Science, 2021, 48 (02): 87-92.
- [13] Tian Yuan, Yuan Ye, Liu Hai Bin, Man Zhi Bo, Mao Cun Li, BERT pre-trained language model for defective text classification of power grid equipment [J]. Journal of Nanjing University of Science and Technology, 2020, 44 (04): 446-453.
- [14] Zhao Yan, Li Xiao Hui, Zhou Yun Cheng, Zhang Yue. A study on agricultural text classification method based on naive bayesian [J]. Water Saving Irrigation, 2018(02):98-102.
- [15] Chen Ping, Kuang Yao, Hu Jing Yi, Wang Xiang yang, Cai Jing. Text categorization method with enhanced domain features in power audit field [J]. Journal of Computer Applications, 2020, 40 (S1): 109-112.
- [16] Liu Zi Quan, Wang Hui Fang, Cao Jing, Qiu Jian. A classification model of power equipment defect texts based on convolutional neural network [J]. Power System Technology, 2018, 42 (02): 644-651.
- [17] Ge Xiao Wei, Li Kai Xia, Chen Ming, Text classification of nursing adverse events based on CNN-SVM [J]. Computer Engineering & Science, 2020, 42 (01): 161-166.
- [18] Wang Meng Xuan, Zhang Sheng, Wang Yue, Lei Ting, Du Wen, Research and application of improved CRNN model in classification of alarm texts [J]. Journal of Applied Sciences, 2020, 38 (03): 388-400.
- [19] Wang Si Di, Hu Guang Wei, Yang Si Yu, Shi Yun, Automatic transferring government website e-mails based on text classification [J]. Data Analysis and Knowledge Discovery, 2020, 4 (06): 51-59.
- [20] Zhang Bo, Sun Yi, Li Meng Ying, Zheng Fu Qi, Zhang Yi Jia, et al. Medical text classification based on transfer learning and deep learning [J]. Journal of Shanxi University (Natural Science Edition), 2020, 43 (04): 947-954.

Identifying Influential Nodes with Centrality Indices Combinations using Symbolic Regressions

Mohd Fariduddin Mukhtar¹, Zuraida Abal Abas², Amir Hamzah Abdul Rasib³
Siti Haryanti Hairol Anuar⁴, Nurul Hafizah Mohd Zaki⁵, Ahmad Fadzli Nizam Abdul Rahman⁶
Zaheera Zainal Abidin⁷, Abdul Samad Shibghatullah⁸

Faculty of Communication and Information Technology, Universiti Teknikal Malaysia Melaka, Melaka, Malaysia^{1,2,4,5,6,7}
Faculty of Mechanical and Manufacturing Engineering Technology, Universiti Teknikal Malaysia Melaka, Melaka, Malaysia^{1,3}
Institute of Computer Science & Digital Innovation, UCSI University, 56000 Cheras, Kuala Lumpur, Malaysia⁸

Abstract—Numerous strategies for determining the most influential nodes in a connected network have been developed. The use of centrality indices in a network allows the identification of the most important nodes in the network. Specific indices, on the other hand, cannot search for a network's entire meaning because they are only interested in a single attribute. Researchers frequently overlook an index's characteristics in favour of focusing on its application. The purpose of this research is to integrate selected centrality indices classified by their various properties. A symbolic regression approach was used to find meaningful mathematical expressions for this combination of indices. When the efficacy of the combined indices is compared to other methods, the combined indices react similarly and outperform the previous method. Using this adaptive technique, network researchers can now identify the most influential network nodes.

Keywords—Centrality indices; combination; symbolic regressions; influential nodes

I. INTRODUCTION

For years, the prediction of the most influential nodes has been a source of contention [1]. The node with the most impact is ranked first, and the one with the least effect is ranked last [2]–[4]. Several research had been carried out to enlist the importance of nodes detection which is such as in finding importance suppliers [5], [6], detection of cancer or virus gene [7] or as well as to monitor the terrorist activities [8].

Over the years, over 403 indices have emerged from the four major indices: Degree Centrality (DC), Betweenness Centrality (BC), Eigenvector Centrality (EV), and Closeness Centrality (CC). Individual task and node importance priorities are claimed to impact the development of various indices. Various centrality measures have been employed to predict node outcomes, with the underlying assumption being that the more centrally situated the nodes are in the network, the greater their spreading potential [9], [10].

However, there were limitations to using indices as a single centrality metric because they could focus on one application area. DC for example, is a good indicator of a node's total connections [11]. Still, it does not necessarily imply a node's value in linking nodes or how central it is to the main group. CC on the other hand, determines how close a node is, but an independent network will not profit from its supremacy if two nodes are placed in distinct components [9]. In the case of BC,

the result will be zero if many nodes are not on the shortest possible path to the remainder of the network [12].

A single centrality metric proved to be insufficient for accurately predicting the network's most important nodes [13]. Combining centrality indices to determine the most influential nodes has been floated. According to [7], [14], there is no single centrality measure that can accurately identify key nodes, but the combination of at least two centrality measures is the most accurate. Combining multiple indices is considerably more accurate than using one index while assessing a node's influence capacity [11]. The influence of a node may be evaluated by its location and surroundings.

The researchers have recursively investigated this topic and determined that each indices have distinct features. This attribute, known as network topology, is reflected differently by different methods, and the evaluation results may include flaws or deficiencies. Borgatti [15] observed two forms of network topology: geodesic paths and walking paths. In the following research, Ashtiani [16] determined that centrality measures may be categorized into five classes based on the reasoning and formulas used. This characterization of centrality indices is also used by [17] while accessing the topological structure of student network.

The goal of this research is to investigate the effectiveness of centrality indices combination. Genetic programming-based symbolic regression (SR) is used here to find expressions to combine the selected indices. Two datasets examined to test the performance of a mix of indices based on their individual properties. Vignery's topology principles guided the selection of the centrality indices in this study. Combination outcomes are compared to results from a previous combination strategy to get a better sense of how effective the combinations are. At the end of this research, the possibility on applying symbolic regression to combine centrality indices will be clarified, and whether the categorization of indices according to their characteristics similarities has an impact on the combined indices.

II. MATERIALS AND METHODS

A. Data

Zachary and Les Misérables (Les-M) datasets, both weighted and unweighted, were used in this investigation. Thirty-four people from the karate club were included in the

Zachary dataset, which documented 78 connections between members who interacted outside the club. The novel Les-M features 77 nodes and 254 edges, including co-occurring characters. Both networks were depicted in the Fig. 1, with information on the most connected nodes.

B. Theoretical Topology of Centrality Indices

The definition of Vignery's eleven centrality indices is simplified and shown in Table I. Several of the indices had the same features and hierarchical clustering analysis is executed to observe whether the indices can be clustered into a single component to justify where the indices are converging. The dendrogram is built up by clustering observations and their similarity levels at each stage and assessing the similarity (or distance) levels of the produced clusters. As a first stage in the modeling technique, the value for each index is computed for each node. Following that, the indices was categorized based on their commonalities. The higher a cluster's similarity level, the more related the variables in that cluster are. These indices are divided into five theoretical groups, explained in Table I.

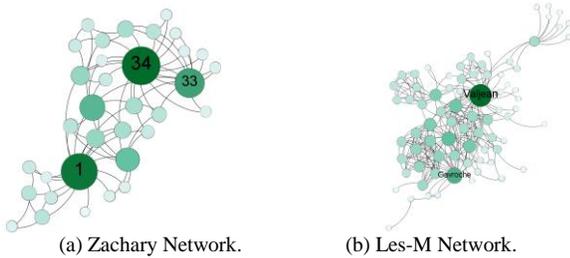


Fig. 1. Representation of both Networks.

TABLE I. VIGNERY'S CENTRALITY INDICES SELECTION

Type of centrality	Definition	Indices involved
Geodesic Distance and Path	A simple measure of the shortest path between two vertices. The length of a geodesic path is called geodesic distance or shortest distance.	Eccentricity (EC), Closeness (CC), Residual closeness (RC), Geodesic K-path (GK), Betweenness (BC), Bottleneck (BN)
Connectivity	Number of direct connections a node has, which means that every pair of vertices has a path linking them.	Eigenvector (EV), Hub & Authority (HUB & AUT), PageRank (PR)
Neighbor's prestige	Measures the number of shortest pathways a specific node has between any other two nodes and is based on the premise that a node with the shortest paths controls communication flows.	EV, HUB & AUT, PR
Nodes' adjacent	The neighbourhood of a node is the collection of all vertices near the vertex.	Cross-clique connectivity (CQ), Maximum Neighborhood Component (MNC)

C. Combinations of Indices

Network dynamics are examined as a function of the structure. The best estimate is made by combining a given number of indices based on the features of the component clusters. Genetic programming (GP) with symbolic regression (SR) is employed to generate mathematical expressions that may predict the simulation response values based on the topological indices used. SR is a technique that uses collected data to construct mathematical equations that may be used to test hypotheses [18]. With SR, the parameters and equation form are automatically searched, unlike typical regression methods that require a fixed-form model built from prior knowledge. GP is commonly utilized in SR because of the high computational complexity imposed by a vast search space that generates new solutions using the notion of biological evolution as a meta-heuristic [18], [19]. SR method's results will then be compared with those from other methods that use centrality indices as a comparison metric. Three algorithms are chosen for comparison which is provided by Eq. (1), (2) and (3).

- C(v) algorithm: Wang [20] developed a combination formula with the integration of DC, diffusion degree (DD) and BC as denoted in Eq. 1 with $a + b + c = 1$.

$$C(v) = aDC(v) + bDD(v) + cBC(v) \tag{1}$$

- BC and Katz (BKC) algorithm: Zhang [21] merged BC and Katz's centrality. Eq. 2 expresses the relationship between BC and KC.

$$BKC(v) = aBC + bKC \tag{2}$$

- Integrated Value of Influence (IVI): Consider topological dimensions [22], IVI consider six key network centrality measurements (normalizes connectivity (NC), ClusterRank (CR), BC, collective influence (CI), DC and local index (LH)) into account as given in Eq. 3.

$$IVI = ((NC' + CR')(BC' + CI'))(DC' + LH'_{index}) \tag{3}$$

III. RESULT AND DISCUSSION

A. Cluster Analysis

Clustering is used to group comparable data objects using a similarity measure. The similarity is a value that displays the strength of a relationship between two data items; it represents how similar data patterns are. The topological framework classifications from Vignery will be extended. We wanted to see how well the indices matched, so we applied a simple hierarchical clustering algorithm.

The results of clustering are depicted in Fig. 2a and 2b. We discovered that for both networks, all the indices could be clustered into four groups. Take note that the component clustering is quite close to Vinery's recommendations. The final partition specifies how the indices will be clustered. In those eleven indices, both networks were supposed to cluster similarly into four groups, except for Les-M, where PR will be in Cluster 2 rather than Cluster 3 (as in Zachary). EC has the

lowest similarity score and is not assigned to any category. EC was not assigned to any cluster because it has the lowest similarity score (10.56 and 7.00) compared to the other indices.

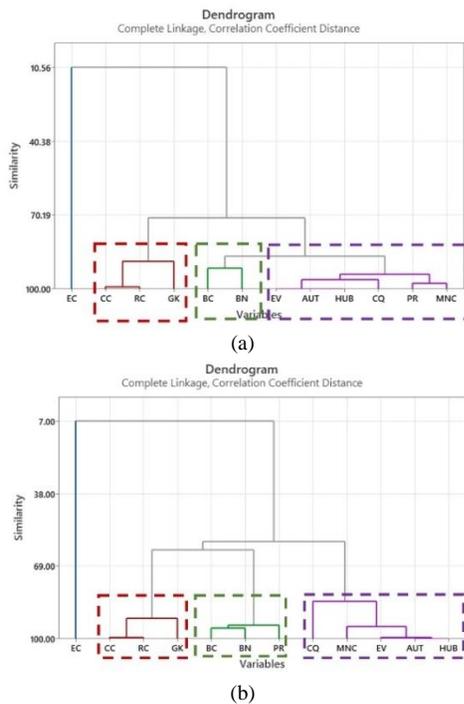


Fig. 2. (a). Hierarchical Clustering of Indices for Zachary. (b). Hierarchical Clustering of Indices for Les-M.

B. Combinations of Indices

Turingbot software was applied to execute the symbolic regression codes, which entails combining a set of base functions into simple formulas to produce a regression model. Fig. 3 depicts the steps we take to generate different mathematical formulations for each cluster. Because the output of SR will vary, we choose the phrase with the lowest value in terms of root mean square error (RMS error) and the highest R-squared (R-sq). Finally, we obtained four distinct expressions for each cluster specified, namely C1, C2, C3, and C. C is an expression that includes all the indices involved. As a result of efficient training and shifting, the analytic equations for both networks are shown in Table II are derived.

C. Pearson Correlation Analysis

The combined indices and component clusters employed in the earlier approach are compared. The correlation technique can be used to discover the similarity of combined centrality indices. The dataset also includes an average value (AVE) for the average result for each node from the Combined, IVI, BKC, and C(v). This AVE value will serve as the reference result to which the correlation converges.

Correlation for both networks show a significant and favorable relationship as shown in Tables IIIa and IIIb. In Zachary, there is a high correlation between AVE and IVI, BKC, and C(v). There is a high association between IVI and the clustered group and all other combined indices. The concept of combining indices while considering their spreaders and hubs can be extended for future use. With a correlation coefficient of 0.797, the relationship with C is likewise satisfactory. It is fascinating to notice that the C1 and C2 are more closely linked to AVE than is C, while C3 is obviously diverged from correlation toward others.

Results for Les-M also give results like Zachary's network when looking at each cluster component, with C1 and C2 being more correlated than C and C3. However, it is interesting to observe that IVI is quite diverging for this network. According to these findings, the SR modeling combination model has successfully identified influential nodes. In the following section, the node's ranking position is being observed.

D. Node's Ranking of Position

In this section, the placements of nodes were analyzed and arranged in descending order. When comparing procedures side by side in Tables IVa and IVb, the top ten ranking position of each approach is considered. Node AVE is a reference column that contains the average positioning value of nodes for each method, as expressed by the average positional value of nodes for each technique. Zachary and Les-M discovered that BKC and C(v) have a very similar node detection to AVE when comparing the two algorithms.

Nodes in IVI react similarly in Zachary, whereas they deviate significantly in Les-M. However, when compared to C for both networks, C1 and C2 show more comparable node detection for both networks when compared to C. To better understand the similarity result, we use the Jaccard similarity (JS) score and Kendall's tau-b to access the top ten nodes' ranking positions. Kendall's tau-b indices are used to measure the strength of a method's ranking position to understand the similarity result from the JS-score better.

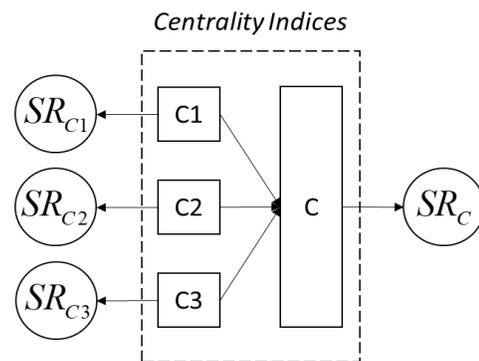


Fig. 3. Illustrations on Indices Combination for each Group.

TABLE II. MATHEMATICAL EXPRESSION OF COMBINED INDICES

Zachary	Les-M
Cluster 1 (C1): CC RC GK $0.00103544 \left(1.33583 + \left(0.915295 RC - \left((-0.122916) (GK - \cos(GK)) \right) \right) \right)$ RMS Error = 0.000248416 R-Sq = 0.987144	Cluster 1 (C1): CC RC GK $(-0.00727966) / \left(0.0565855 * \left(0.802185 * \left(0.121368 * "GK" + "RC" \right) \right) - 2.51961 \right)$ RMS Error = 0.000248416 R-Sq = 0.987144
Cluster 2 (C2): BC BN $\left(2.12409 - \tanh \left(\text{round} \left(-0.457495 + \tan(BN) \right) + \tan(0.208093BN) \right) \right) \times$ $\left(\text{round} \left(-2.28163 - \cos(BN) \right) + \left(\left(\cos(-0.390946 - \text{floor}(0.542588BN)) \right) / 0.316869 + BN \right) \right)$ RMS Error = 0.841825 R-sq = 0.998876	Cluster 2 (C2): BC BN PR $(-3.95423e-05) \times \left(\tan(0.00897528 + (BN - 1.00369 * BC)) + 169.68 \right) \times$ $\left(\cos \left((-0.329319) * \left(\cos(BN) + BC \right) \right) - (368.183 + BC) \right)$ RMS error = 0.00283 R-sq = 0.946660
Cluster 3 (C3): EV AUT HUB PR CQ MNC $(-0.00344533 + PR) \times$ $\left(\cos \left(\left(0.41957 + MNC - (CQ - (MNC / 0.39926)) \right) \right) \times AUT \right) + 0.066212 + 2.55262 * HUB$ RMS Error = 0.0356482 R-Sq = 0.985527	Cluster 3 (C3): EV AUT HUB CQ MNC $\log 2(CQ) + \left(\left(3.98809 + \tan(1.13497 - (-0.0407873 + (-0.0729471 + CQ))) \right) \right)$ $\left(- \left(AUT / \tan \left(\tan \left(\tan(0.0705538 * (-0.0449964 + CQ)) \right) \right) \right) \right)$ $\left(\times (-0.092534 + AUT) \right)$ RMS Error = 0.2957 R-Sq = 0.99513
Combine (C): EC CC RC GK BC BN EV AUT HUB PR CQ MNC $0.000275397 \times \left(GK - \left(0.0113009 * \left(\left(\tan(1.23832 - GK) - \cos(MNC) + \left(-7.32811 + 0.979185 + MNC - \left(\tan(RC - 0.814871) + (BN - \tan(BC)) \right) \right) \right) * BN \right) \right) \right)$ $\left(-(\sinh(HUB) - RC - 9.00789) \right)$ RMS Error = 0.008983 R-Sq = 0.97642	Combine(C): EC CC RC GK BC BN EV AUT HUB PR CQ MNC $\left(("PR" / (0.0140288 + 0.000763124 * "BC" - "EV")) + 4.15684 \right)$ $\times (1.32335 * "MNC" + "RC" - 9.22003) - (-103.165 * "HUB")$ $+ \cos("EC" - 8.01939 * "RC")$ RMS = 5.4994 R-Sq = 0.9941

TABLE III. (A) CORRELATIONS IN ZACHARY

	IVI	BKC	C(v)	C	C1	C2	C3
BKC	0.910						
C(v)	0.823	0.929					
C	0.712	0.736	0.605				
C1	0.752	0.726	0.592	0.975			
C2	0.800	0.856	0.735	0.778	0.762		
C3	0.484	0.525	0.439	0.278	0.268	0.366	
AVE	0.951	0.988	0.943	0.724	0.729	0.839	0.510

TABLE III. (B). CORRELATIONS IN LES-M

	IVI	BKC	C(v)	C	C1	C2	C3
BKC	0.320						
C(v)	0.539	0.943					
C	0.718	0.641	0.852				
C1	0.445	0.691	0.841	0.872			
C2	0.487	0.925	0.954	0.770	0.785		
C3	0.050	-0.004	0.042	0.079	0.087	0.093	
AVE	0.570	0.944	0.997	0.851	0.817	0.953	0.031

TABLE IV. (A) TOP TEN NODES BASED ON RANKING IN ZACHARY

Rank	Node AVE	Node IVI	Node BKC	Node Cv	Node C	Node C1	Node C2	Node C3
1	1	1	1	1	1	1	1	34
2	34	34	34	34	3	34	3	22
3	33	33	33	33	34	3	33	1
4	3	2	3	28	32	33	32	3
5	32	3	32	3	14	32	34	2
6	2	4	9	32	33	2	24	25
7	14	8	2	6	9	9	9	15
8	9	24	14	23	20	14	2	5
9	6	9	20	27	2	20	14	18
10	24	14	6	26	4	4	20	17

TABLE IV. (B) TOP TEN NODES BASED ON RANKING IN LES-M

Rank	Node AVE	Node IVI	Node BKC	Node Cv	Node C	Node C1	Node C2	Node C3
1	Valjean	Gavroche	Valjean	Valjean	Valjean	Valjean	Valjean	Simplice
2	Gavroche	Courfeyra	Myriel	Gavroche	Gavroche	Marius	Myriel	Toussaint
3	Marius	Bahorel	Gavroche	Marius	Marius	Gavroche	Gavroche	Woman2
4	Fantine	Joly	Marius	Fantine	Javert	Javert	Fantine	Tholomye
5	Myriel	Combefer	Fantine	Myriel	Enjolras	Thenardie	Marius	Bamatabo
6	Thenard	Feuilly	Thenardie	Thenardie	Thenardie	Enjolras	Cosette	Mlle.Gille
7	Enjolras	Valjean	Javert	Javert	Bossuet	Bossuet	Enjolras	Gillenorm
8	Javert	Enjolras	Mlle.Gille	Enjolras	Courfeyra	Cosette	Montparn	Fantine
9	Bossuet	Mabeuf	Enjolras	Bossuet	Bahorel	Fantine	Mme.The	Marius
10	Courfeyr	Grantaire	Tholomye	Mme.The	Joly	Babet	Thenardie	Prouvaire

- Jaccard similarity score

Jaccard similarity (JS) score compares two sets of scores by counting the number of elements in each group. JS can be calculated numerically by dividing the intersection of sets by the union of sets [23]. The higher the value, the greater the correlation between the two data sets. The higher the Jaccard similarity indices, the closer two sets of data are to one. Definition of JS is formulated as in Eq. 4.

$$JS(A, B) = \frac{\text{number of observations in both sets}}{\text{number in either set}} = \frac{|A \cap B|}{|A \cup B|} \quad (4)$$

For Zachary, C has a higher similarity score than Les-M when looking at the Jaccard scores for the two networks. If we look at Zachary and compare C to the previous combined technique, we see that C has the same top ten ranking entities for IVI (0.6667), BKC (0.8182), and AVE (0.6667). C has a low degree of similarity (less than 0.5) to IVI, BKC, and AVE in Les-M. C1 and C2 are interestingly comparable to C with the IVI, BKC, and C(v), which follow similar patterns. The JS-score heatmap for both networks are shown in the Tables Va and Vb. The greater the degree of resemblance between methods, the darker is the color.

TABLE V. (A) JACCARD SIMILARITY SCORE FOR ZACHARY

	C1	C2	C3	C	IVI	BKC	C(v)	AVE
C1	1							
C2	0.8182	1						
C3	0.25	0.25	1					
C	0.6667	0.8182	0.25	1				
IVI	0.6667	0.6667	0.25	0.6667	1			
BKC	0.8182	0.8182	0.25	0.8182	0.5385	1		
C(v)	0.3333	0.3333	0.1765	0.3333	0.25	0.4286	1	
AVE	0.6667	0.8182	0.25	0.6667	0.667	0.8182	0.4246	1

TABLE V. (B) JACCARD SIMILARITY SCORE FOR LES-M

	C1	C2	C3	C	BKC	AVE	C(v)	AVE
C1	1							
C2	0.5385	1						
C3	0	0.1111	1					
C	0.5385	0.3333	0	1				
IVI	0.1765	0.1765	0	0.4286	1			
BKC	0.5385	0.5385	0.25	0.4286	0.1765	1		
C(v)	0.6667	0.6667	0.4286	0.5385	0.1764	0.6667	1	
AVE	0.5385	0.4286	0.4286	0.4286	0.1765	0.5386	0.6667	1

• Kendall's tau-b

When comparing the rankings of different methods, Kendall's tau-b is applied to determine the ordinal relationship between pairs of observations [24], [25]. Correlation strength and direction are measured using Kendall's tau-b correlation coefficient. According to the theory of rank correlations, the closer two sets of data are linked together, the more closely they are related. Within this range, positive and negative numbers can indicate concordance or discordance, which is characterized by increasing or decreasing values, respectively. The correlation value between two variables increases when the ranks of the observations are similar; the correlation value decreases when the positions of the observations are different. Kendall's tau-b is defined as in Eq. 5.

$$\tau_b(A, B) = \frac{m_c - m_d}{\sqrt{(m - T_0)(m - T_1)}} \quad (5)$$

The top ten ranking of nodes using Kendall's tau-b results for Zachary and Les-M are shown in Tables VIa and VIb. Comparing C with IVI, BKC, C(v), and AVE shows that C has a rather low and moderate positive tau value in the Zachary network, while it has a high-rank similarity for the Les-M network. C also has significance and a strong positive tau correlation with C1 (0.644). C1 also shows the importance and positive correlation with C2. Observe from AVE analysis shows that results were significant in Les-M compared to Zachary except for C3. Since there were differences in C2 and C3 for Zachary and Les-M, it might affect the way on the rank behavior.

TABLE VI. (A) KENDALL'S CORRELATION SCORE FOR ZACHARY

	C1	C2	C3	C	IVI	BKC	C(v)	AVE
C1	1							
C2	0.067	1						
C3	-0.244	-0.467	1					
C	-0.022	0.467	-0.566*	1				
IVI	0.111	-0.111	-0.156	0.067	1			
BKC	0.378	0.244	-0.244	0.156	0.556*	1		
C(v)	0.2	0.244	0.022	0.333	0.467	0.467	1	
AVE	0.244	0.2	-0.378	0.022	0.6*	0.6	0.156	1

*. Correlation is significant at the 0.05 level (2-tailed).

TABLE VI. (B) KENDALL'S CORRELATION SCORE FOR LES-M

	C1	C2	C3	C	IVI	BKC	C(v)	AVE
C1	1							
C2	0.556*	1						
C3	0.244	-0.2	1					
C	0.644*	0.467	0.156	1				
IVI	0.111	0.111	0.067	0.378	1			
BKC	0.644*	0.644*	0.067	0.644*	0.289	1		
C(v)	0.556	0.467	0.156	0.822*	0.467	0.733*	1	
AVE	0.6*	0.511*	0.111	0.867*	0.511*	0.778*	0.956*	1

*. Correlation is significant at the 0.05 level (2-tailed).

IV. CONCLUSION

The selection of influential nodes is crucial for fostering knowledge and behavior adoption in a network because they can influence other nodes. It is possible to gain a better understanding of network structure and behavior by using prominent nodes. The importance of centrality in identifying influential network spreaders in this scenario cannot be overstated.

Our results show that combining centrality indices can identify the influential nodes in a network. To combine these indices, symbolic regression can identify appropriate mathematical expressions that will fit the network's features. When it comes to recognizing significant nodes, the newly constructed mathematical expression's function performs similarly or better than previous methods (IVI, BKC, and C(v)) that were validated using Pearson correlation, Jaccard similarity score, and Kendall's tau-b correlation of ranking.

It was discovered while clustering indices based on similar attributes that each cluster component may have the same impact as aggregating all indices. Clustering can reduce the total number of indices to be combined while achieving the same overall result. It was also discovered that index selection is critical. A few indices, including Katz centrality and Closeness centrality, could not be computed. Katz's centrality fails to detect connections between high-centrality nodes. To function, closeness centrality requires a well-connected network, and it fails when two nodes belong to different components. Large, complex datasets necessitate more computationally intensive methods. Clustering for the indices involved would be difficult, as it is here.

In future work, the selection of suitable indices to be combine is important. Analyzing their features as well as the computational time required to run each of the indices may be put into factor selection. It is also a good idea to run another run-up for a different network with a different number of weighted nodes if possible.

ACKNOWLEDGMENT

The author would like to extend a special thank you to the FRGS-RACER/2019/FTKMP-COSSID/F00412 grant, Centre of Research and Innovation Management (CRIM), Faculty of Mechanical and Manufacturing Technology, Universiti Teknikal Malaysia Melaka, Malaysia for funding this research. Appreciation also goes for Faculty of Communication and Information Technology and Ministry of Education Malaysia.

REFERENCES

- [1] Z. A. Abas et al., "Analytics : a Review of Current Trends , Future," *CompuSoft*, vol. 9, no. 1, 2020.
- [2] Y. Yang, L. Yu, X. Wang, S. Chen, Y. Chen, and Y. Zhou, "A novel method to identify influential nodes in complex networks," *International Journal of Modern Physics C*, vol. 31, no. 2, Feb. 2020, doi: 10.1142/S0129183120500229.
- [3] A. Zareie, A. Sheikhhahmadi, M. Jalili, M. Sajjad, and K. Fasaie, "Finding influential nodes in social networks based on neighborhood correlation coefficient" vol. 194, p. 105580, 2020, doi: 10.1016/j.knosys.
- [4] M. Alshahrani, Z. Fuxi, A. Sameh, S. Mekouar, and S. Huang, "Efficient algorithms based on centrality measures for identification of top-K influential users in social networks," *Information Sciences*, vol. 527, pp. 88–107, Jul. 2020, doi: 10.1016/j.ins.2020.03.060.

- [5] N. J. Pulles, J. Veldman, and H. Schiele, "Identifying innovative suppliers in business networks: An empirical study," *Industrial Marketing Management*, vol. 43, no. 3, pp. 409–418, 2014, doi: 10.1016/j.indmarman.2013.12.009.
- [6] B. B. M. Shao, Z. (Michael) Shi, T. Y. Choi, and S. Chae, "A data-analytics approach to identifying hidden critical suppliers in supply networks: Development of nexus supplier index," *Decision Support Systems*, vol. 114, pp. 37–48, Oct. 2018, doi: 10.1016/j.dss.2018.08.008.
- [7] M. Jalili et al., "Evolution of centrality measurements for the detection of essential proteins in biological networks," *Frontiers in Physiology*, vol. 7, no. AUG. Frontiers Media S.A., Aug. 26, 2016. doi: 10.3389/fphys.2016.00375.
- [8] K. Basu, C. Zhou, A. Sen, and V. H. Goliber, "A novel graph analytic approach to monitor terrorist networks," *Proceedings - 16th IEEE International Symposium on Parallel and Distributed Processing with Applications, 17th IEEE International Conference on Ubiquitous Computing and Communications, 8th IEEE International Conference on Big Data and Cloud Computing*, 11t, pp. 1159–1166, 2019, doi: 10.1109/BDCloud.2018.00171.
- [9] Q. Ma and J. Ma, "Identifying and ranking influential spreaders in complex networks with consideration of spreading probability," *Physica A: Statistical Mechanics and its Applications*, vol. 465, pp. 312–330, Jan. 2017, doi: 10.1016/j.physa.2016.08.041.
- [10] H. Zhou, M. Ruan, C. Zhu, V. C. M. Leung, S. Xu, and C. M. Huang, "A Time-Ordered Aggregation Model-Based Centrality Metric for Mobile Social Networks," *IEEE Access*, vol. 6, pp. 25588–25599, Apr. 2018, doi: 10.1109/ACCESS.2018.2831247.
- [11] A. Ibnoulouafi, M. el Haziti, and H. Cherifi, "M-Centrality: Identifying key nodes based on global position and local degree variation," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2018, no. 7, Jul. 2018, doi: 10.1088/1742-5468/aace08.
- [12] J. Wang, C. Li, and C. Xia, "Improved centrality indicators to characterize the nodal spreading capability in complex networks," *Applied Mathematics and Computation*, vol. 334, pp. 388–400, Oct. 2018, doi: 10.1016/j.amc.2018.04.028.
- [13] K. Raman, N. Damaraju, and G. K. Joshi, "The organisational structure of protein networks: Revisiting the centrality-lethality hypothesis," *Systems and Synthetic Biology*, vol. 8, no. 1, 2014, doi: 10.1007/s11693-013-9123-5.
- [14] G. Scardoni, G. Tosadori, M. Faizan, F. Spoto, F. Fabbri, and C. Laudanna, "Biological network analysis with CentiScaPe: centralities and experimental dataset integration," *F1000Res*, vol. 3, no. 0, p. 139, 2014, doi: 10.12688/f1000research.4477.1.
- [15] S. P. Borgatti, "Centrality and network flow," *Social Networks*, vol. 27, no. 1, pp. 55–71, 2005, doi: 10.1016/j.socnet.2004.11.008.
- [16] M. Ashtiani et al., "A systematic survey of centrality measures for protein-protein interaction networks," *BMC Systems Biology*, vol. 12, no. 1, pp. 1–17, 2018, doi: 10.1186/s12918-018-0598-2.
- [17] K. Vignery and W. Laurier, "A methodology and theoretical taxonomy for centrality measures: What are the best centrality indicators for student networks?," vol. 15, no. 12 December. 2020. doi: 10.1371/journal.pone.0244377.
- [18] B. K. Petersen, T. N. Mundhenk, S. K. Kim, C. P. Santiago, and J. T. Kim, "Deep symbolic regression," 2021.
- [19] S. Kim et al., "Integration of Neural Network-Based Symbolic Regression in Deep Learning for Scientific Discovery," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 9, 2021, doi: 10.1109/TNNLS.2020.3017010.
- [20] W. Jianwei, R. Lili, and G. Tianzhu, "A new measure of node importance in complex networks with tunable parameters," *2008 International Conference on Wireless Communications, Networking and Mobile Computing, WiCOM 2008*, pp. 1–4, 2008, doi: 10.1109/WiCom.2008.1170.
- [21] Y. Zhang, Y. Bao, S. Zhao, J. Chen, and J. Tang, "Identifying Node Importance by Combining Betweenness Centrality and Katz Centrality," in *Proceedings - 2015 International Conference on Cloud Computing and Big Data, CCBBD 2015*, Apr. 2016, pp. 354–357. doi: 10.1109/CCBD.2015.19.

- [22] A. Salavaty, M. Ramialison, and P. D. Currie, "Integrated Value of Influence: An Integrative Method for the Identification of the Most Influential Nodes within Networks," *Patterns*, vol. 1, no. 5, Aug. 2020, doi: 10.1016/j.patter.2020.100052.
- [23] M. R. Hamedani and S.-W. Kim, "JacSim: An accurate and efficient link-based similarity measure in graphs," *Information Sciences*, vol. 414, pp. 203–224, Nov. 2017, doi: 10.1016/j.ins.2017.06.005.
- [24] D. A. Walker, "JMASM9: Converting Kendall's Tau For Correlational Or Meta-Analytic Analyses," 2003.
- [25] Y. Y. Keng, K. H. Kwa, and C. McClain, "Convex combinations of centrality measures," *Journal of Mathematical Sociology*, vol. 45, no. 4, pp. 195–222, 2021, doi: 10.1080/0022250X.2020.1765776.

Improving Computational Thinking in Nursing Students through Learning Computer Programming

Leticia Laura-Ochoa, Norka Bedregal-Alpaca, Elizabeth Vidal
Universidad Nacional de San Agustín de Arequipa
Arequipa, Peru

Abstract—Computational thinking is a fundamental skill for problem-solving, it uses computational concepts and other types of thinking such as algorithmic. The experience of improving computational thinking in nursing students using block-based programming environments such as Code.org, Lightbot, and the Python textual programming language is described. The results obtained are analyzed by applying a pre-and post-test of computational thinking to the students. The methodological design is quasi-experimental since it did not work with a control group. The study group was made up of 30 students from the Professional School of Nursing of the National University of San Agustín de Arequipa. The results show that teaching programming allows the understanding of computational concepts and improves computational thinking. It is concluded that block-based programming environments and the Python language facilitate the development of algorithmic thinking and computational thinking.

Keywords—Computational thinking; computational thinking assessment; computational thinking test; programming; programming environments

I. INTRODUCTION

Computational Thinking (CT) is a fundamental skill that influences different disciplines and professions, not only those related to science and engineering [1][2]. It is considered a transversal competence that goes beyond the use of computers and coding [3] since it includes algorithmic and parallel thinking, which involve different types of thought processes, such as compositional reasoning, pattern matching, procedural thinking, and recursive thinking. [2], which are required by new generations of students in different areas of knowledge.

According to Román-González et al. [4], programming is the main demonstration of the ability of the CT through the use of the computer, because it allows the development of algorithmic thinking, problem-solving, logic, and debugging skills [5]. In this context, textual programming is considered the final educational goal at the end of the K-12 level in most countries [4], because students in adolescence value the higher cognitive load of textual languages. However, text-based programming can make learning difficult for beginning students, overwhelming their cognitive ability [6]. Furthermore, it is considered a difficult task due to the lack of complete development of computational thinking in students [7]. Therefore, it is necessary to select the appropriate didactic tools and strategies that facilitate the teaching of programming with an approach oriented to the development of computational thinking. For [8], block-based programming environments are a good way to introduce beginning students to programming.

In addition, in [9] they consider that these environments facilitate the understanding of programming concepts; but both block-based and text-based programming environments allow the development of skills related to computational thinking.

In addition, Tikva and Tambouris [10] have found that teachers face challenges in incorporating TC practices, so more capacity building frameworks and interventions that support teachers for successful integration of TC into their teaching practices are required. They believe that the relationship between tools and TC development should be explored as future work to provide information on which tools support which TC learning strategies.

Consequently, the aim of this work is to show the programming tools that were used in the experience and how they favor the acquisition of computational concepts and the development of computational thinking practices and skills, which can be of help and reference for teachers who need to incorporate computational thinking in their teaching work. This experience was carried out in an online learning environment with higher education students from the professional school of nursing, a career different from science and engineering, where the majority of students are usually women.

In this work, the experience of the use of programming environments based on blocks and text is described, to improve the development of computational thinking skills for beginning students in programming a Basic Computer course, which includes as a learning unit the Introduction to programming. An analysis of the pre- and post-test results of computational thinking applied to students is carried out to verify the improvement of computational thinking. In the experience, Code.org and Lightbot block programming environments were used prior to the Python textual programming language to facilitate the understanding of computational concepts and the acquisition of computational thinking practices.

II. RELATED WORK

In the work of Brackmann et. al. [11], a quasi-experiment was presented in two primary schools in Spain, to develop the computational thinking skills of students through disconnected activities, students' ages are between 10 and 12 years old. Their results show that the students of the experimental group, who participated in the disconnected activities, significantly increased their computational thinking skills, compared to those who did not participate in the disconnected activities, evidencing the effectiveness of the disconnected approach for the development of computational thinking skills. However,

they consider that this approach has limitations, and there may come a point where it loses its effectiveness and the use of computing devices is required to further develop these skills.

Vasquez and Luján [12] carried out an evaluation of aptitude level on the development of computational thinking in students of the basic level of secondary school. They identified the need to strengthen skills related to computational thinking such as analysis, algorithm design, and data abstraction, because their average scores did not exceed 50% of the total number of questions evaluated. Likewise, they found that the maturity of the students and the cognitive development according to the academic degree do not establish the level of development of computational thinking skills. Since their cultural environment must also be considered. The authors considered that the results of the computational thinking test can be used to design and develop a computational thinking course that allows strengthening skills that require it. Montes-León et al. [7] describe their experience of the application of computational thinking activities that positively influences the improvement of learning in a course of fundamentals of programming. They carried out an analysis of the results of a pre and post test of computational thinking applied to secondary students divided into control and experimental groups, to evaluate the improvement of computational thinking. The ages of the participating students were between 15 and 16 years old. They also analyzed the results of the evaluation applied in the course. The activities that were carried out in the experimental group were some exercises from the international Bebras contest, mathematical problems, exercises from a university entrance test and mental games.

Laura-Ochoa and Bedregal-Alpaca [13] found that the incorporation of computational thinking practices allowed students to improve their performance on the first Python programming course, where they used support tools such as PSeInt, CodingBat and the turtle graphic library for the development of skills related to computational thinking, conducted an analysis of the grades obtained by the students of the control and experimental group in the midterms and final average of the programming course, where the students of the experimental group showed an improvement in their learning results. As future work, they suggest the application of computational thinking measurement assessments by following a practice-oriented programming teaching approach to computational thinking.

III. METHODOLOGY

The methodological design used was quasi-experimental, since it did not work with a control group. In the study group, there were 30 students enrolled in the Basic Computer course (groups B and F) of the academic semester 2020-B of the Professional School of Nursing of the National University of San Agustín de Arequipa (Peru).

The Basic Computer Science course at the Professional School of Nursing of the National University of San Agustín de Arequipa - Peru, is given in the second academic semester. It is developed for 17 weeks. It has 4 practice hours a week, lasting 50 minutes each, equivalent to 2 credits.

The students of the experiment were 30 women (100%), who participated in the pre- and post-test of computational thinking, as well as in the development of the third learning unit: Introduction to Programming. This unit is developed during five weeks, with two weekly sessions of 2 hours each.

In the practice hours, the students experimented with the use of visual programming environments: Code.org, Lightbot and Python textual programming language. The method used in the class sessions was expository-participatory.

The computational thinking test developed by Román-González et al. [14] was used to measure the level of development of computational thinking in students. Its test is consistent with other computational thinking tests aimed at middle/high school students [15], it is mainly aimed at Spanish students between 12 and 14 years old (7th and 8th grade of primary school), but it can be used in lower and higher grades. It is aligned with some CT computational concepts defined by Brennan and Resnick [16] and partially aligned with some computational practices.

The students who participated in the experience completed the computational thinking test [14] in the first week of classes on the subject and at the end of the third learning unit: Introduction to programming, accessing an online test through a Google Apps form.

To measure the improvement of computational thinking, a comparison of the scores obtained by the students in the pre-test and post-test of computational thinking was made, to check if the activities carried out in the third learning unit: Introduction to Programming allowed the acquisition of computational concepts and development of skills related to computational thinking.

IV. DESCRIPTION OF THE EXPERIENCE

Block-based visual programming environments (Code.org, Lightbot) and Python programming language were selected for teaching the introduction to programming and development of skills related to computational thinking in the third learning unit of the Basic Computer course taught to students of the nursing professional career in the period 2020-B.

Table I shows the programming topics that were developed using the selected programming environments, where the students learned computational thinking concepts considered in the work of Luo, Antonenko and Davis [17], such as sequential instructions, conditionals, and loops.

Students began learning the block-based visual programming language using Code.org by completing exercises in the "hour of code" tutorials that enabled them to understand concepts of sequential statements, loops, and functions. An example of block programming on Code.org is shown in Fig. 1, where students understood the utility of iterative statements by replacing repetitive blocks of code (left) with loops (right) to draw the geometric figure of the square, in which the students acquired some computational thinking practices such as iteration and abstraction through the identification of repetitive patterns, which allowed them to acquire the ability to reduce unnecessary details and propose new solutions.

TABLE I. PROGRAMMING TOPICS

Tools	Sequential Instructions	Conditional Instructions	Repetitive Instructions	Functions
Code.org	X		X	X
Lightbot	X		X	X
Python	X	X	X	

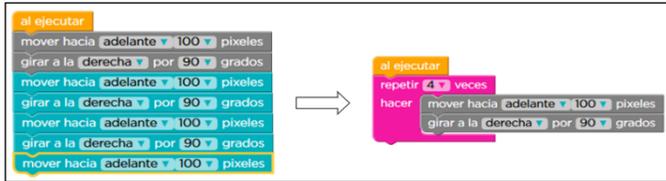


Fig. 1. Replacing Repetitive Blocks of Code with Loops at Code.org.

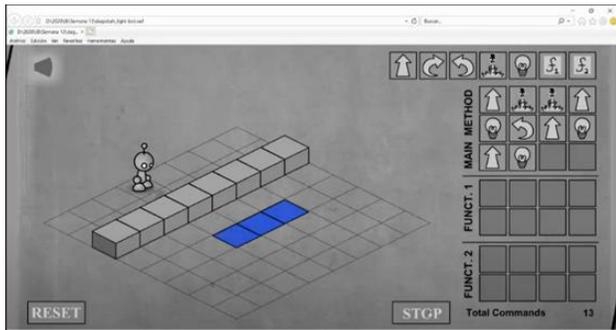


Fig. 2. Sequential Instructions with Repeating Patterns.

The Lightbot video game was used to reinforce algorithmic thinking skills in students by thinking in sequence of instructions for problem solving. Fig. 2 shows an example of the Lightbot third level solution using only sequential instructions (move, jump, turn left) in the MAIN METHOD for the robot to move and turn on all blue blocks, which is the objective of the video game, where repetitive patterns such as moving forward and turning on are identified.

In Fig. 3, a second solution for the third level of Lightbot is shown, where students, through generalization (identification of repetitive patterns go forward and turn on), abstraction (reduction of unnecessary details) and decomposition of the program using one of the functions (FUNCT. 1), acquire the ability to find better solutions to the problem and make use of code reuse.

In addition, the students practiced the Lightbot video game from Code.org, where they used recursion to create loops in the PROC1 procedure (Fig. 4).

After experience with the block-based programming environments, the students learned the syntax and semantics of the Python textual programming language using the Google Collaboratory environment for the creation and execution of code, with which they reinforced their understanding of computational concepts, such as sequential statements, conditionals, and loops. Fig. 5 shows some examples of codification of single and double selection structures; but there was also done double selection instructions (nested).

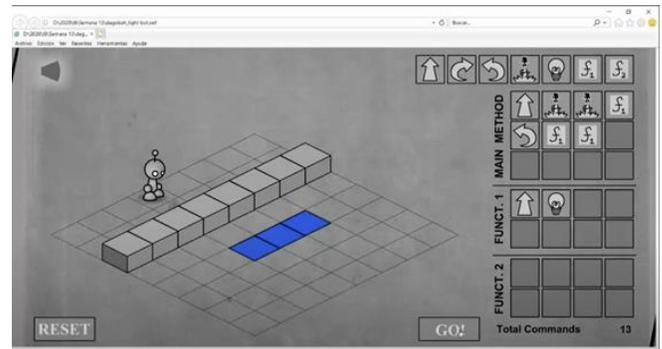


Fig. 3. Program Decomposition and Code Reuse with Functions.

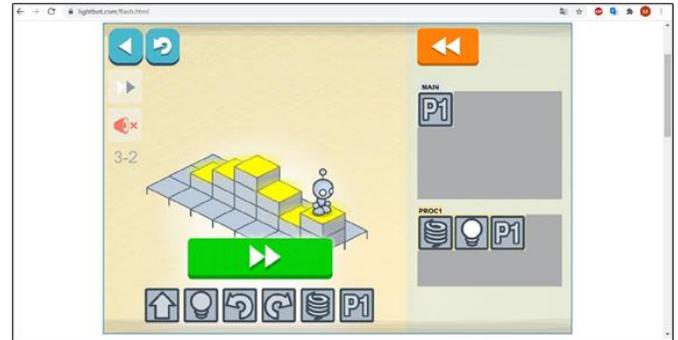


Fig. 4. Looping using Recursive Calls in Lightbot from Code.org.

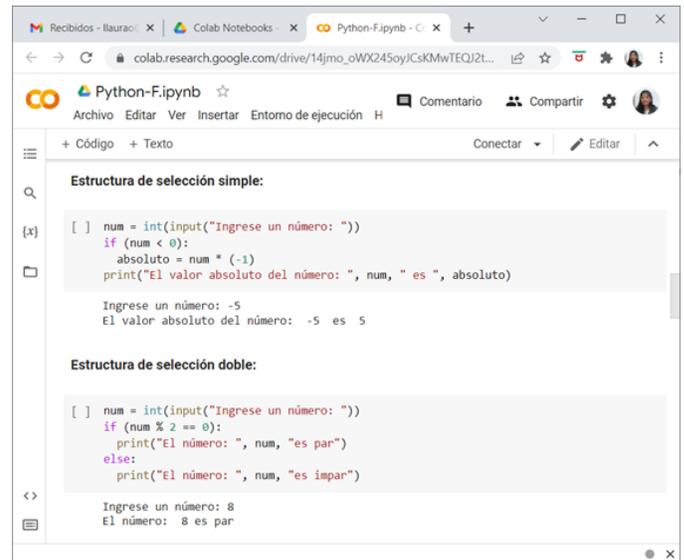


Fig. 5. Coding with Single and Double Select Structures using Python.

With the programming environments used in the experience, computational thinking skills are used, such as automation and debugging for the execution of programs and logical analysis for the verification of the results.

Therefore, students acquired computational thinking practices such as abstraction, decomposition, iteration, and debugging.

V. RESULT AND DISCUSSION

The computational thinking test [14] was applied at the beginning of the Basic Computer Science course before taking Unit 3 of Introduction to programming, serving as a pre-test for the evaluation of said unit.

Fig. 6 shows the percentage of correct answers per question in the pre-test. The regression line shows the progressive difficulty of the test. The average correct percentage of the 28 questions was 64.64%.

At the end of the third learning unit, the computational thinking test was applied again [14].

Fig. 7 shows the percentage of correct answers per question in the post-test. The regression line shows an improvement in terms of the progressive difficulty of the pre-test. The average correct percentage of the 28 questions was 80.6%.

Table II shows some descriptive statistical data related to the total scores obtained by the students in the pre and post-test. The total scores are evaluated from 0 to 28.

Fig. 8 and 9 show histograms with the distribution of said total scores, where the improvement in the total, mean, median and mode scores of the post-test are evidenced. In the post-test, the median and mode have the same value of 23.

Fig. 10 shows box plots for the scores obtained through computational thinking pre- and post-test. In the post-test, an atypical value is observed that corresponds to a student who obtained a total score of 15 points. The median, maximum value, minimum value, quartiles are higher in the post-test.

Table III shows the averages of the percentages of correct answers of the questions for the computational concepts that are addressed in the computational thinking test, obtained by the students in the pre and post-test.

TABLE II. DESCRIPTIVE STATISTICAL DATA OF THE TOTAL SCORES

	Pre-Test	Post-Test
Minimum	11	15
Maximum	26	27
Mean	18.1	22.567
Median	17.5	23.0
Mode	16	23
Standard deviation	3.791	2.885

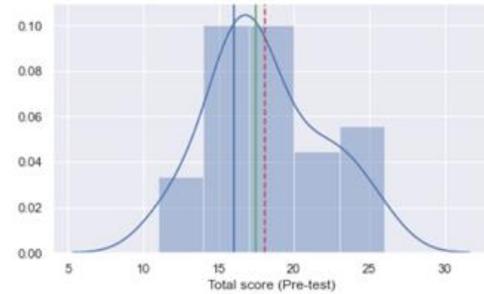


Fig. 8. Histogram with the Distribution of Total Scores (Pre-test).

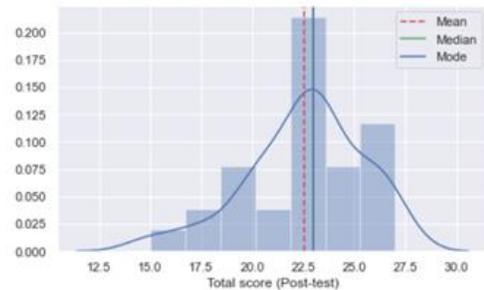


Fig. 9. Histogram with the Distribution of Total Scores (Post-test).

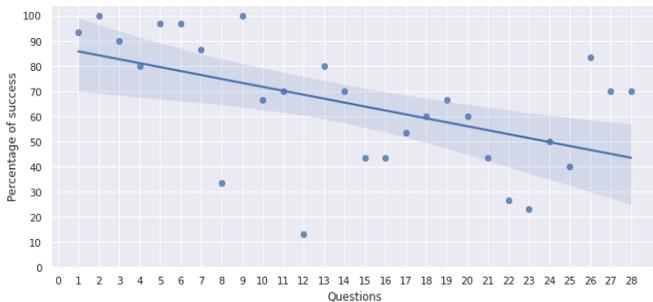


Fig. 6. Success Percentage per Question in the Pre-test.

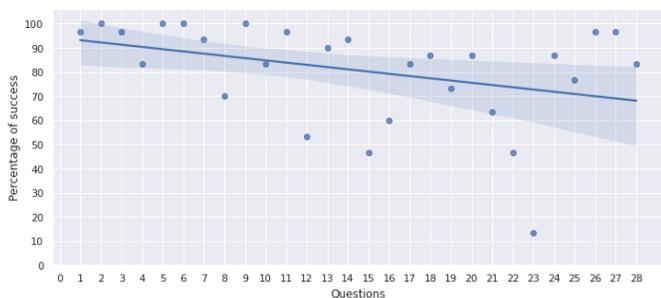


Fig. 7. Success Percentage per Question in the Post-test.

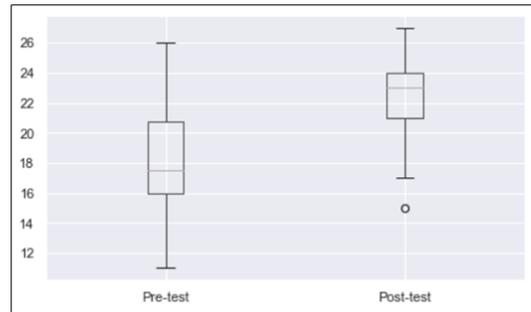


Fig. 10. Box Plots of the Scores Obtained in the Pre and Post Test.

TABLE III. SUCCESS PERCENTAGES FOR COMPUTATIONAL CONCEPTS

	Pre-Test (%)	Post-Test (%)
Basic directions and sequences	91	94
Loops 'Repeat Times'	78	91
Loops 'Repeat Until'	63	83
Simple Conditional 'if'	59	73
Complex conditional 'if/else'	60	83
While conditional	36	53
Simple functions	66	88

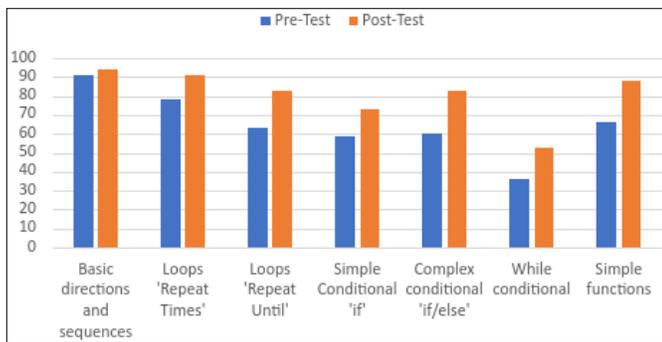


Fig. 11. Success Percentage for Computational Concepts.

Fig. 11 shows that the percentages of success in the post test are higher for each of the computational concepts addressed in the test.

Therefore, it can be affirmed that the performance of the students has improved in the post-test, evidencing an improvement in computational thinking, after the development of the third learning unit, in which the introduction to programming was made using visual programming environments based on Code.org blocks, Lightbot and the Python textual programming language.

According to the experience described, the programming environments used allowed the acquisition of computational practices such as abstraction, decomposition, iteration and debugging, which correspond to the computational thinking practices defined by Brennan and Resnick [16] and adapted in the work of Luo, Antonenko and Davis [17]. In addition, students developed skills such as algorithmic thinking, generalization, automation, and debugging, which are part of the computational thinking skills identified in five featured articles in the work of [18]. It was achieved that students had a means to acquire methodological tools, deepen knowledge, and cultivate skills [19].

We agree with a previous work [20] that learning text-based programming is important for the development and professional performance of students, because in addition to reinforcing concepts and practices of computational thinking, they can expand their learning for the creation of applications, analysis or visualization of data.

In programming courses aimed at beginners, we consider it important to carry out activities related to computational thinking at the beginning of the course to improve their learning, where visual programming environments based on blocks can be used. Likewise, in [8] they consider block-based programming environments a good way to introduce beginning students to programming.

Likewise, we agree with Zapata-Cáceres et al. [21], in that computational thinking is not limited only to the activities of computer scientists; it is applied in daily life and in different areas of knowledge, so it is a necessary skill to adapt to the future.

VI. CONCLUSION

This article has presented the experience of developing computational thinking skills through a block and text-based

programming activities to improve students' computational thinking. We examined the total scores obtained by the students in the pre-test and post-test, there is evidence of an improvement in the scores in the post-test with the programming environments used in the learning unit of introductory programming, which indicates the effectiveness of programming for understanding computational concepts addressed in the test. We concluded that the visual programming environments based on blocks in combination with the textual programming language using Python allow the student to acquire computational concepts and computational thinking practices such as abstraction, decomposition, iteration, and debugging in introductory programming courses directed mainly to beginning students of non-computer related careers.

ACKNOWLEDGMENT

The authors' thanks are expressed to the National University of San Agustín de Arequipa for the support received in the realization of the proposal and the results are expected to benefit the institution.

REFERENCES

- [1] J. M. Wing, "Computational thinking," *Communications of the ACM*, vol. 49, no. 3, pp. 33-35, 2006.
- [2] J. M. Wing, "Computational thinking: What and why. The Link," *News from the School of Computer Science at Carnegie Mellon University*, 2011.
- [3] J. Acevedo-Borrega, J. Valverde-Berrosco and M. D. C. Garrido-Arroyo, "Computational Thinking and Educational Technology: A Scoping Review of the Literature," *Education Sciences*, vol. 12, no. 1, pp. 39, 2022.
- [4] M. Román-González, J. C. Pérez-González, J. Moreno-León and G. Robles, "Can computational talent be detected? Predictive validity of the Computational Thinking Test," *International Journal of Child-Computer Interaction*, vol. 18, pp. 47-58, 2018.
- [5] F. Buitrago Flórez, R. Casallas, M. Hernández, A. Reyes, S. Restrepo and G. Danies, "Changing a generation's way of thinking: Teaching computational thinking through programming," *Review of Educational Research*, vol. 87, no. 4, pp. 834-860, 2017.
- [6] C. Chen, P. Haduong, K. Brennan, G. Sonnert and P. Sadler, "The effects of first programming language on college students' computing attitude and achievement: a comparison of graphical and textual languages," *Computer Science Education*, vol. 29, no. 1, pp. 23-48, 2019.
- [7] H. Montes-León, R. Hijón-Neira, D. Pérez-Marín and S. R. Montes-León, "Mejora del Pensamiento Computacional en Estudiantes de Secundaria con Tareas Unplugged," *Education in the knowledge society (EKS)*, no. 21, pp. 24, 2020.
- [8] Z. Xu, A. D. Ritzhaupt, F. Tian and K. Umaphy, "Block-based versus text-based programming environments on novice student learning outcomes: a meta-analysis study," *Computer Science Education*, vol. 29, no. 2-3, pp. 177-204, 2019.
- [9] L. Laura-Ochoa and N. Bedregal-Alpaca, "Análisis de entornos de programación para el desarrollo de habilidades del pensamiento computacional y enseñanza de programación a principiantes," *Revista Ibérica de Sistemas e Tecnologías de Informação*, no. E43, pp. 533-548, 2021.
- [10] C. Tikva and E. Tambouris, "Mapping computational thinking through programming in K-12 education: A conceptual model based on a systematic literature Review," *Computers & Education*, vol. 162, pp. 104083, 2021.
- [11] C. P. Brackmann, M. Román-González, G. Robles, J. Moreno-León, A. Casali and D. Barone, "Development of computational thinking skills through unplugged activities in primary school," in *Proceedings of the 12th workshop on primary and secondary computing education*, pp. 65-72, 2017.

- [12] A. J. O. Vasquez and B. I. S. Luján, "Evaluación del nivel de aptitud desarrollo de pensamiento computacional en jóvenes de nivel básico secundaria," *RECIE. Revista Electrónica Científica de Investigación Educativa*, vol. 4, no. 2, pp. 1151-1163, 2019.
- [13] L. Laura-Ochoa and N. Bedregal-Alpaca, "Incorporation of Computational Thinking Practices to Enhance Learning in a Programming Course," *International Journal of Advanced Computer Science and Applications(IJACSA)*, vol. 13, no. 2, 2022, DOI: 10.14569/IJACSA.2022.0130224.
- [14] M. Román-González, J. C. Pérez-González and C. Jiménez-Fernández, "Test de Pensamiento Computacional: diseño y psicometría general," in *III congreso internacional sobre aprendizaje, innovación y competitividad (CINAIC 2015)*, pp. 1-6, 2015.
- [15] M. Román-González, J. C. Pérez-González and C. Jiménez-Fernández, "Which cognitive abilities underlie computational thinking? Criterion validity of the Computational Thinking Test," *Computers in human behavior*, vol. 72, pp. 678-691, 2017.
- [16] K. Brennan and M. Resnick, "New frameworks for studying and assessing the development of computational thinking," in *Proceedings of the 2012 annual meeting of the American Educational Research Association*, 2012.
- [17] F. Luo, P. D. Antonenko and E. C. Davis, "Exploring the evolution of two girls' conceptions and practices in computational thinking in science," *Computers & Education*, vol. 146, pp. 103759, 2020.
- [18] S. Bocconi, A. Chiocciariello, G. Dettori, A. Ferrari and K. Engelhardt, "Developing computational thinking in compulsory education - Implications for policy and practice," in *JRC Science for Policy Report*, 2016.
- [19] N. Bedregal-Alpaca, "Virtual tutoring and blended-learning in the postgraduate course: Orientations and results of an experience," *Proceedings of the 17 LACCEI international Multi-conference for Engineering, Education and Technology*, 2019, DOI 10.18687/LACCEI2019.1.1.220.
- [20] L. Laura-Ochoa and N. Bedregal-Alpaca, "Development of Computational Thinking Skills: An Experience With Undergraduate Students," in *2021 XVI Latin American Conference on Learning Technologies (LACLO)*, IEEE, pp. 112-117, 2021, DOI 10.1109/LACLO54177.2021.00070.
- [21] M. Zapata-Cáceres, E. Martín-Barroso and M. Román-González, "Computational thinking test for beginners: Design and content validation," in *2020 IEEE Global Engineering Education Conference (EDUCON)*, IEEE, pp. 1905-1914, 2020.

Improving Social Engineering Awareness, Training and Education (SEATE) using a Behavioral Change Model

Azaabi Cletus, Benjamin Weyory, PhD, Alex Opoku, PhD
School of Sciences, University of Energy and Natural Resources, Sunyani, Ghana

Abstract—Social Engineering (SE) Awareness, Training, and Education (SEATE) is one of the recommended defenses against SE attacks among users of Information Systems. However, many of these SEATE programs fails to achieve the desired impact leading to exposures. This study sought to explore SEATE programs to identify gaps/challenges and propose relevant content, Delivery Methods, and a novel behavioral change Model to improve SEATE programs among users. An explorative Literature Search was conducted on the relevant SEATE Content, Delivery methods and the challenges of SEATE Programs. Consequently, the relevant and critical content and delivery methods were proposed. The challenges that impede the efficient and effective conduct of SEATE Programs were established. A behavioral change Model known as Social Engineering Awareness, Transition, Adaptation and Consolidation (ATAC) based on Stable-Quasi-Stationary Equilibrium theory was proposed. The model was validated using Expert Opinions. Five (5) expert in cybersecurity were recruited to appraise the model based on five metrics; fit for purpose, novelty, ease of use and structure. The results show that, challenges still exist in the conduct of SEATE programs. To improve SEATE programs requires relevant and innovative content, and delivery method (Hybrid Approach). Validation of the proposed behavioral change model showed an average score at 73.6% and performance metrics at 92%. As the menace of SE attacks rages on and exploiting the user, the need for SEATE programs remains imperative. A well-developed and relevant content, delivery methods and a clear understanding of the challenges is required to improve SEATE. Following the model developed, and the repeated use of it will lead to improving user resistance and or immunity to SE attacks and by extension improve security culture among users.

Keywords—Social engineering; user training; user awareness; user education; ATAC model

I. INTRODUCTION

Globally, cyber-attacks remains a major threat affecting individuals, small and medium enterprises, multi-national corporations, nation states and indeed all global stakeholders in the cyber space [1], [2], [3], [4], [5]. This is occasioned by the ushering in of the 4th Revolution (information superhighway), the growth and expansion of the internet, the Internet of Things (IOT), cloud computing and extensive penetration of smart phone telephony [6].

Even though these statistics are positive signals towards cyber inclusion, the problem of ensuring that, the data and information stored in computers and in Critical Information

Infrastructure(CII) are protected against unauthorized access, modification, vandalism and others poses a big challenge particularly attacks against the human wall (the weakest link) also known as social engineering (SE). [7] describe, SE as gaining access to systems, buildings, data by exploiting humans using psychology instead of using technical procedures to break in. [8] sees it as influencing a person to an action that may or may not be in his/her interest. Consequently, the increased use and adoption of these technological assets has expanded the cyber-attack surface in general and SE in particular, resulting in exposures to critical information asset and the concomitant effect of reputational loss, financial loss, legal issues [9].

Consequently, cyber criminals realizing that, the ‘wet ware’ is easier to compromise have resorted to employing SE attack methods to perpetuate cybercrimes by gaining access to confidential information [10]. Hence, the need for programs to protect users against such SE attacks.

Over the years, defenses against social Engineering attacks have been varied. The most common SE defenses has been user education [11] [12], user Awareness [9] [5], user Training [13]. Other recent defense programs included Gamification [14]), the use of Apps [15], Serious games [16], Virtual labs [17], conferences and tournaments [5]. The rest of the programs include the use of predictive and preventive tools [18] and Recognition tools [19].

As organization realizes the impact of user awareness as a means to complement the technology-based defenses, many have increased budgetary allocation, time and effort to ensure security among users using policies and other behaviour-based approaches such as awareness, training and education [20], [21].

Even though these programs are aimed at building the resistance of users and ensuring that they are well prepared to defend against various SE attacks, they fail to achieve their intended purpose due to how these programs are organized, the content and the pedagogy used; thus, leading to exposures of confidential information and its consequential impact.

This paper sought to explore SEATE programs establishing the challenges, exploring and proposing relevant Content, delivery methods, and a model to be used to conduct SEATE effectively and efficiently that will lead to permanent SE security culture, resistance to SE attacks and reflective

behavior of users when faced with an SE attack. To do this, the following research objectives were set:

- Explore and proposed relevant SE Content, Delivery methods and challenges of SEATE programs.
- To propose a behavioral change model to improve SEATE programs among users.

The contribution/value/novelty of this work;

There is a limited academic study on the use of behavioral change models in improving SEATE programs. Consequently, this study and its findings is a modest contribution to SEATE programs in particular and improvement of user resistance in general. Hence, this study is a modest contribution to knowledge in the field of SE in particular and Cybersecurity in general.

Specifically, the study contributed to knowledge in the following ways:

1) Critically analyzed literature and established relevant content, Delivery methods and challenges of SEATE. Through this approach, we proposed innovative and relevant SEATE Content and the key points that should be included and emphasized during SEATE Programs.

2) We also highlighted the industry delivery methods and their challenges and thus proposed a hybrid approach so as to complement the deficiencies in each of them.

3) Proposed a model to be followed to improve SEATE resulting in improvement of 92% in model performance metrics rating.

4) Contributed in design process, methodology that can be used by practitioners to improve upon their SEATE projects.

The rest of the paper is structured as follows:

In Section 2, Theoretical and Related Works; Section 3, Content, Delivery Methods and challenges of SEATE Programs; Section 4, Proposed Model and Validation, Section five 5, Results, section 6, Discussions of the Findings, Section 7, Conclusion and Future Works and at the end are the references of the study.

II. THEORETICAL FRAMEWORK AND RELATED WORK

A. Theoretical Framework

The concept of SE mainly refers to attacks aimed at tricking the user (Holder of a vital information Asset) to divulge such information against the wish of the user [8]. As an attack against the user, any defense or protective mechanism should aim at the user. This will ensure that, the user is aware of such attacks, modify their behavior about SE attacks and manage the needed change to prevent, and or mitigate the attack.

The study [22] is regarded as the father of change management (CM). He proposed the 3 –step model indicating that, a successful change passes through 3 steps; unfreezing, moving and refreezing [22]. He contended that to manage change process, the organization must unfreeze; change from

current state to a neutral position, to enable the unlearning of the old behavior, and to ensure that the new behavior can be adopted and adapted successfully. Once the change occurs, the organization refreezes into the new state. This is often referred to as Stable –Quasi- stationary- Equilibrium.

Extending and applying this theory, we indicated, that, SE as a cyber-phenomenon, requires that all stakeholders are offered the required SEATE with the aim of improving resistance to such attacks, creating permanent cyber/SE security culture and consciousness and permanent behavior change against SE attacks.

Reasoning on this principle, we proposed a model known as Awareness, Transition, Adaptation and Consolidation (ATAC) model to improve SEATE programs. A review of related works in social engineering awareness, trainings programs follows in the next section.

B. Related Work

Research into security Awareness program in general and social engineering in particular has gained pace in recent years especially in programs aimed at improving security against SE attacks [5], [9], [21],[6].

The author in [6], proposed an educational model for systematic adaptation to cyber security training programs. However, this model is for generic cyber security awareness and fails to address the issue of SE. Specifically, using the modus operandi in social engineering differs with other technology or traditional hacking methods.

A web-Based System (SAWIT tool kit) was proposed and translated into a prototype to improve security awareness. It was based on knowledge sharing among employees [21].

The author in [5] delivered a conference paper on challenges of implementing training and awareness programs targeting cybersecurity social engineering. They suggested budgetary constraints for trainings, lack of understanding of information security, bad organizational cyber security culture as some of the challenges facing SEATE. He recommended the use of security preparedness exercises and awareness programs as a means to improve security.

The author in [23] proposed a framework to evaluate the risk inherent in the Internet of Things (IOTs) based on the situational awareness. The focused on awareness in IOT devices and how promote situational awareness of security. Other studies considered SE awareness on the bases of the business environment such as technology, organization etc. as a way to improve SEATE. Social issues as a limitation against SEATE was also conducted [11].

Notwithstanding the number of studies conducted in social engineering awareness trainings, not much is done in clearly identifying the key critical challenges of a SEATE programs, the required and relevant content and delivery methods and a model that can improve user behavior change to ensure permanent cybersecurity culture in general and SE in particular. Thus this paper explored relevant SEATE programs, content, delivery methods and the challenges and proposed a novel behavioral change model for the improvement of SEATE programs. The next section explored

the Content, Delivery methodologies and challenges/setbacks of SEATE programs.

III. CONTENT, DELIVERY METHODS AND CHALLENGES OF SEATE PROGRAMS

A. Cybersecurity Social Engineering Content

Social Engineering attackers continue to plague the cyber world with new and novel attacks. Even though organization is spending huge sums of money to ensure security, their effort always most times fails due to user vulnerability to social engineering attacks. To ensure improved security, organizations' employees' knowledge need to be improved using awareness, training and education programs [20]. This should include exposure to security policies, processes and best standards that promote corporate cybersecurity in general and social engineering in particular. The content should contain awareness programs at the base for starters. These are programs aimed at exposing the user to SE attacks towards changing the behaviour of the user [2].

Training programs should follow this, which enables the user to make appropriate security choices in their daily personal and work life. Users are trained on specific actions to take in specific cases and should be selected and implemented based on the set objectives.

Finally, the SEATE program should graduate to education where individuals interested to take careers in cybersecurity are given specialized education in specific area by providing in-depth knowledge in the area of security. Thus, SEATE programs should follow a learning continuum, which begins with awareness creation, cumulatively building into training and eventually evolves into education as shown in Fig. 1 known as the cybersecurity learning continuum.

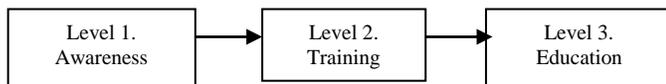


Fig. 1. Learning Continuum.

The author in [20] suggests that most information security awareness programs are generic in nature, too much information leading to information overload. This makes it difficult for users to decipher the relevant content to concentrate. Even with the relevant content, the delivery approach is also relevant to ensure that the right and relevant content is well delivery to the user.

B. SEATE Program Delivery Methods

There are many and varied methods used in the delivery of SEATE programs. [2] Suggested face-face method (lectures, storytelling and workshops), self-directed learning which can be static in nature (text and web based) or flexible/dynamic (videos and games) and finally, teachable options such as embedded delivery methods such as online learning. In the face-face, approach involves a physical environment with or without an expert who facilitates the process. The self-directed learning involves a virtual platform where the SEATE program is delivered such as web-based trainings, text-based and video based approaches. In the case of teachable delivery

method, embedded links and content is attached for user to learn.

[24] contend that, security awareness delivery methods include the conventional methods such as posters, stickers, leaflets, newsletters; Instructor- led; formal presentations, training sessions and online delivery methods such as electronic articles or emails, web-based security awareness methods, alert messages and game-based methods.

It is worthy of note that even though many of these approaches are proposed towards achieving maximum benefit from such programs, many fail due to inherent challenges such as too much information, cost, boring, inexperience instructors, monotony leading to security breaches.

C. Challenges to Cybersecurity SEATE Programs

The use of SE security training awareness, education and other programs aimed at protecting users against SE attacks is well documented in literature [25],[9]. These programs aim at improving user resistance, and increased SE attack consciousness. [5]opine that, several factors militate against SE training and awareness; the Business Environment, Social issues including industry competition, the compliance or legal frameworks with the country, organizational issues, economic and personality issues or traits that serves as challenges to a successful SE education, training and awareness.

Other methods to SE education, training and awareness include the current methods such as games, Apps, Virtual labs, tournaments, conferences [5]. However, these have challenges such as coordination in the case of serious games, and personality issues about collaborative approaches. Others include real-life simulations and videos. These simulations are generic and fail to cater for the individuals in the organization.

The earlier methods to SEATE programs include manual reminders, the use of posters, awareness campaigns, online courses and physical access programs [20]. However, these methods are said to be boring, tedious and time consuming and lack practical exposure for employees.

From the forgoing, time constraints, Budgetary constraints, generalized nature of the training and awareness program without recourse to the individual users, characteristics such as educational level, organization level (operational, management and levels) pose a major challenge in the successful conduct of SEATE projects. Thus, to be able to effectively and efficiently carry out SEATE programs to achieve the intended objective, there is the need for relevant content, innovative delivery methods and organizational and behavioral change models [21]. The next section will consider the characteristics of relevant content, innovative delivery method and propose a social engineering security behavior change model for effective and efficient SEATE programs.

IV. METHODOLOGY

The study aimed at exploring the SEATE content, delivery methods and the challenges faced in achieving the intended objectives and to propose innovative content and delivery methods and a novel behavioral change model to improve SEATE programs.

TABLE I. PROPOSED RELEVANT SEATE CONTENT AREA AND DESCRIPTION

Comprehensive Knowledge of a Social Engineer.	Clear understanding of the goals, objectives and motives of a Social Engineer, types, characteristics and tricks
Comprehensive knowledge of vectors in use.	Understanding of both semantic, syntactic and AI based vectors, forms of vectors, categories, their deployment strategies and how to overcome them.
Comprehensive knowledge of users/ victims.	Understanding of user vulnerabilities, traits that makes users vulnerable, level of training and exposure to cybersecurity issues etc.
PsychoSocial factors used in social engineering attacks.	Clear understanding of the psychosocial factors used in carrying out an attack; strong effect, diffusion or responsibility, overloading authority, urgency etc.
Relevant standards and regulations.	ISO/IEC27001 & 27002, PCI/DSS, FISMA, GRAMM-LEACH BLILEY ACT, HIPAA, Red Flag Rule, GDPR. These will provide users and third party contractors with policies, procedures, cardholders information, information assets risk management, responsibilities and compliance, employee management and training, system failures, create awareness to raise red flag as when a threat shows up, general data protection to monitor, compliance, awareness, training and audit to ensure data security.

To achieve this, literature was explored to establish the challenges of SEATE programs. Then a meta-data analysis of the industry-based SEATE Contents was explored and compared with our proposed innovative content as illustrated in Table I.

Secondly, we compared the traditional delivery methods, identified their gaps, and proposed an innovative Hybrid method for delivering SEATE programs. The proposed hybrid approach is shown in Fig. 2. The use of the hybrid was proposed because the weaknesses in the traditional methods will be complemented by combining them.

To improve overall SEATE programs objectives, we proposed a behavioral change model known as Awareness, Transition, Adaptation and Consolidation (ATAC) to improve SEATE programs. The figure below shows the model and brief description of it.

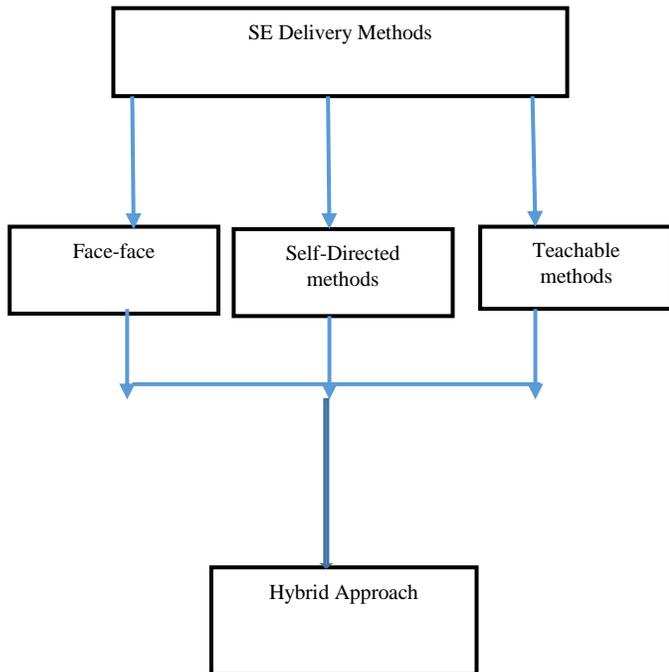


Fig. 2. Proposed Hybrid Delivery Approach.

A. Proposed Model and Description

We proposed a model known as “Awareness, Transition, Adaptation and Consolidation (ATAC) as a solution to improving SEATE and by extension S.E Attacks. The Model is as shown in Fig. 3.

The model is in four (4) phases; Awareness, Transition, Adaptation and Consolidation with an arrow depicting the needed force of change such as cybersecurity consciousness (internal or external). This is how the model works:

1) *Awareness phase:* Making users aware of their behavior deficiency to SE risks is needed. When users’ awareness of the implication of their current deficient state is made known and the associated Danger, they begin to think of how to change.

2) *Transition state:* Users are now aware of the dangers associated with their behavior; Hence, wanting to change from their current state to the proposed new state.

3) *Adaptation phase:* In this stage, the user has transitioned into the new or required state and are now prepared to live such a new life; being security conscious and taking calculated actions to ensure that his behaviour does not lead to exposure.

4) *The consolidation phases:* this phase ensure that, the use is now ready and living the desire organization security culture. Enforcing this new behaviour through monitoring, reminders, penetration testing will ensure that the user do not relapse to the old state.

Such a cycle of creating awareness among users and exposing them to the dangers leads them to want to change. This leads to transition where the user switches from the lack of knowledge to the new state. This desire enables the user to adapt to the new way of cybersecurity conscious life. Consistent use of this will lead to consolidation, where the desired behavior is enforced to become part of the user and the process continues back to the awareness when new requirement for change is necessitated. To ensure the potential usefulness and usability of the proposed model, it needs to be validated.

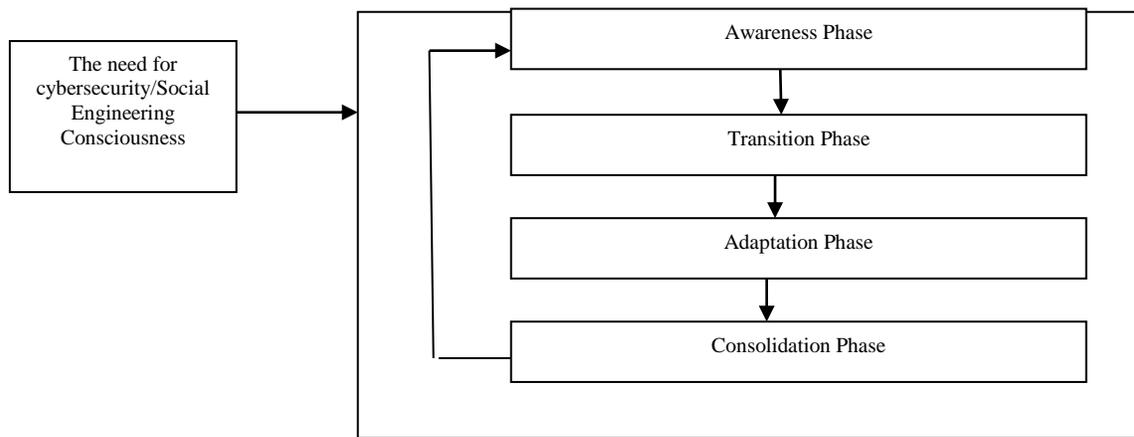


Fig. 3. Proposed ATAC Model.

TABLE II. COMPARISON OF HYBRID DELIVERY MODEL WITH EXISTING DELIVERY METHODS

Delivery Method	Lecture	Workshop	Story	Embeddded Links	Flexible	Videos Games	Static
Face-Face	✓	✓	✓	X	X	X	X
Self-Directed Teachable Moments	X	X	X	X	✓	✓	✓
Hybrid	X	X	X	✓	X	X	X

B. Model Validation

To validate the model, Expert Opinions were elicited to ascertain the usefulness and usability of the artefact. This was done using Observational Empirical Research whereby the

The researcher did not intervene in the assessment of the model by the Experts. The aim was to gain useful information about the expected usability and usefulness of the proposed artefact in a real-world context [26]. Experts were to rate the model in dimensions/metrics such as fit for purpose, novelty, ease of use, architectural structure. Each metric was to be rated on a scale of 1 to 5 based on the expert’s view of the model to that metric. The result is shown in Table I. Descriptive statistics were used to represent the data. We measured the central tendency using Arithmetic means as this describes the center of the data if divided equally among the subjects (Howard & Fletcher, 2016). The results of the study are presented in the next section.

The result of the proposed relevant SEATE Program Content, the proposed Hybrid delivery method and the Expert Opinion were analyzed and presented as shown in Table I, Fig. 3 and Table II, respectively.

V. RESULT

This study sought to explore SEATE programs and to propose relevant Content, Delivery Methods, Challenges and to propose innovative SEATE Content, Hybrid Delivery method, and a novel behavioral change Model to improve SEATE programs among users. The result of the study was presented according to these objectives using tables and graphs.

In research objective 1, the aim was to propose a relevant SEATE content for the improvement of SEATE programs to obtain the desired impact and results. This is demonstrated in Table I.

The next objective was to compare the existing SEATE delivery methods with our proposed hybrid approach. The result of the proposed SEATE Delivery methods and the proposed innovative approach is as shown in Table II.

To improve the conduct of SEATE programs, we proposed a behavioral change model and evaluated it. The model was validated using Expert opinion as a means to scaling to practice. Experts were to rate the model based on the dimensions given on a scale of 1 to 5 for all the dimensions. The result of the opinions is presented in table. The overall Expert score of the model by Experts is presented in Fig. 4 while that of performance metric measures is shown in Fig. 5.

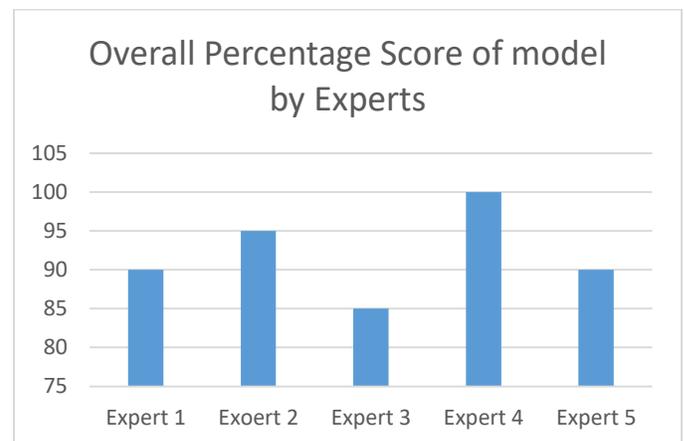


Fig. 4. Overall Percentage Score of the Model by Experts.

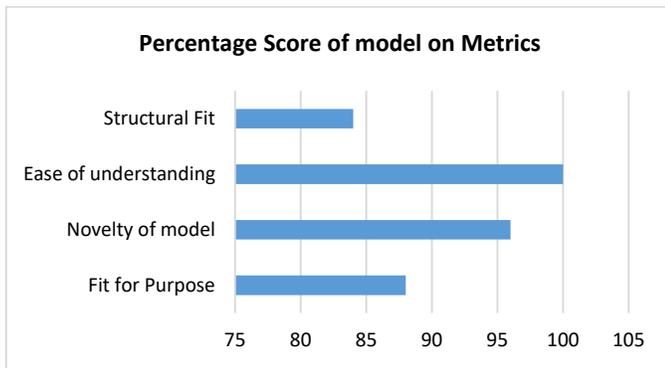


Fig. 5. Percentage Score of the Model on Metrics.

VI. DISCUSSION

The study sought to explore the content, delivery methods, challenges of SEATE programs and suggest Relevant SE content, innovative Delivery method and a model for improving SEATE programs.

The findings from the study demonstrated that, the content of SEATE programs is relevant to the success of the program. To have an effective SEATE Content, a critical analysis of the SE attack cycle is relevant. We argue that, having knowledge of the Social Engineer such as goals/motives(financial, espionage, competitive advantage, revenge), types of social engineers(hackers, penetration testers, disgruntled employees, government's foreign intelligence people, spies) and all relevant information about Social Engineers when included in the SEATE program improves the understanding and knowledge of the user [27]. Also of relevance to improving SEATE Programs content is knowledge of social engineering vectors (syntactic, semantic and AI Based attacks). These vectors include phishing, pharming, water holing, spyware, adware, rootkits, Trojans, etc. [7]. Thirdly, knowledge and comprehensive understanding of the User and the vulnerability to SE attacks is relevant as part of the content to be included in the SEATE programs. Moreover, an effective SEATE program should also include the psychosocial factors used in SE attacks. These among others include, elucidation, strong effect, urgency, reciprocation, diffusion of responsibility, authority [8],[9]. Finally, knowledge of the relevant industry standards and regulations needs to be explained to users. These include ISO/IEC27001 & 27002 to provide all employees, contractors and third parties the policies, procedures, for their job; PCI/DSS for employees to understand the cardholder information and to acknowledge it. Federal information security management Act for employees to understand the information assets, risks, responsibilities and compliance. Gramm-Leach Bliley Act for users to understand the risks to customer information, employee management and training, information systems knowledge, and managing systems failures. Health Insurance Portability and Accountability Act (HIPAA) implement security awareness and training programs for all employees. The use of the Red Flag Rule for user to be able to identify threats and raise Red Flags when there is need and General Data Protection Regulation (GDPR) to create awareness, training, monitor compliance, and audit to ensure Data security as depicted in Table I.

The delivery methods used in conducting SEATE programs has an impact on the outcomes of such programs. Such programs are many and varied. The author in [2] opined that, they are face-face, self-directed and teachable methods. According to [24], the delivery methods include conventional methods, Instructor-led and online methods. However, these methods have their individual downside when used alone. Consequently, we argue that, using all of them together will complement each other's deficiencies. Hence, the proposal for the use of the hybrid approach in delivering based on the peculiarity of the problem being addressed. The use of the hybrid approach will ensure that the weaknesses in each of the proposed traditional methods are complemented and compensated by the other method. This will lead to SEATE programs being effective and leads to improved security to social engineering attacks in particular and cybersecurity in general and SEATE programs in particular.

To propose a behavioural change artefact to improve SEATE programs, we propose a model known as Awareness, Transition, Adaptation and Consolidation (ATAC) following two theories (Conscious Competence Model and the stable Quasi-Random Equilibrium Model): [28],[22].

The CCM is a framework that describes the stages individuals have to pass through when learning a skills or behavior change to move from being unconscious/unskillful to becoming conscious/skillful. It is made up of four stages; Unconscious incompetence, where the individual as unaware, do not understand or know about a particular issue; in this case, SEATE programs and its impact. The second stage is Conscious Incompetence where the individual becomes aware of their skill/knowledge attitudinal deficit, hence, expresses the need to learn the skills or behavior change. At the third stage, called Conscious Competence, the individual have made progress and have acquired a reasonable level of the needed skill, but does things with little difficulty.

Finally, continuous use of the skill knowledge/ attitude over and over leads to Unconscious Competence where he/she performs the art/skill without thinking or with less effort.

The other theory is the stable quasi-stationary equilibrium, which describes change as a transition between two dynamic states in which each of the state itself is dynamic. It comprises unfreeze or unlock from present position to or behavior by creating an enabling environment through education training, motivation. The second is move from the present to the new state by implementing the new way of thinking or attitude. Thirdly; refreeze; by making the new system, the accepted way the system should work [22]. Consequently, following these principles/theories we argue that, the use of the ATAC model not only improve user SEATE knowledge/behavior, but will lead to permanent behavior/knowledge change and lead to permanent immunology among Users against Social Engineering Attacks. An Expert Opinion or Observational Empirical Research conducted showed very high rating for the metrics such as fit for purpose, novelty, ease of use and architecture/structure with 92% on average. This suggests that such a model has the potential to influence behavior change among users with overall average expert score of 73.6%.

VII. CONCLUSION AND FUTURE WORK

The study aimed at Improving Social Engineering Awareness, Training and Education (SEATE) by exploring SEATE Content, Delivery methods, challenges and proposing an innovative Content, a hybrid Delivery method and a behavioral change model as a Non-Technical defense Approach to S E attacks defense.

The findings shows that SEATE fails due to poor content and poor delivery methods coupled with other challenges such as generic nature of training, limited budget provision, poor security policies and compliance and user resistance to changes. The use of relevant an innovative content, hybrid delivery methods, and a behavioral change model improves the conduct of SEATE programs.

As the cyber warfare in general and Social Engineering in particular rages on, with daunting challenges, cyber criminals have found innovative ways of penetrating the parametric defenses and delivering malicious content to users with the aim of compromising their systems. When that happens, the user becomes the last line of defense; to click or not to, update or not to. These critical binary decisions require that users understand the relevant SE Content conducted through well-delivered and innovative delivery methods. Consequently, following the novel model improves the immunity or resistance of users to SE attacks and enables them to know how to react in such attack circumstances. Thus, adopting the novel behavioral change model (ATAC) showed a high potential at improving the conduct of SEATE programs that improved user immunity/Resistance to SE attacks. One limitation of the model is the fact that, it was qualitatively evaluated; future effort will empirically follow the model to conduct a quantitative longitudinal study to practically establish the dimensions of the artefact (model) and to automate same for conduct of SEATE programs.

REFERENCES

- [1] Ahmed Alzarahni. Coronavirus Social Engineering Attacks: Issues and Recommendation. International Journal of Advanced Computer Science and Applications, Volume 11, No.5, 2020.
- [2] Asma A. Alhashmi, Abdulbasit Darem, Jemal H. Abawajy. Taxonomy of Cybersecurity Awareness Delivery Methods: A Countermeasure for Phishing Threats. International Journal of Advanced Computer Science and Applications. Volume 12, No. 10, 2021.
- [3] Symantec Security Summary. Covit 19 Attacks Continuo and New Threats on the rise. Symantec Enterprise Blogs, Retrieved, 20-12-21.
- [4] Global Cybersecurity Index. Measuring Commitment to Cybersecurity. International Telecommunication Union, 2020.
- [5] Hussein Aldawood, Geoffrey Skinner. Reviewing cyber security Social Engineering Training and Awareness Programs-Pitfalls and Ongoing Issues. Future Internet, 2019. MDPI. <https://doi.org/10.3390/fi11030073>.
- [6] George Hatzivasilis, Soltiris Ioannidis, Micheal Smyrlis, George Spanoudakis, Fulvio Frati, Ludger Goeke, Torsten Hildebrandt, George Tsakirakis, Fotis Oikonomou, George Leftheriotis, and Hristo Koshutanski. Mordern Aspert of Cyber-Security Training and Continuous Adaptation of Programmes to Trainees. Applied Sciences, 2020. Doi: 10.3390/app10165702.
- [7] Fatima Salahdine & Naima Kaabouch (2019). Social Engineering Attacks: A survey. Future Internet, MDPI. [doi.10.3390/fi11040089](https://doi.org/10.3390/fi11040089).
- [8] Chritopher Hadnagy. Social Engineering: The Science of Human Hacking, Wiley, Indianapolis. URL: www.wiley.com, 2018.
- [9] Sumar, Musah, Albladi, George, R.S. Weir. Predicting individuals' vulnerability to social engineering in social networks. Cybersecurity. Springer open, 2020. <https://doi.org/10.1186/s42400-02-00047-5>.
- [10] Arroyo, A.M.,Rea,F.;Sandini, G.;Sciutti, A.Trust and social engineering in human robot interaction: will a robot make you disclose sensitive information, conform to its recommendation or gamble? IEEE Robot Autom. Lett.2018,3,3701-3708.
- [11] David Airehrour, Nisha Vasudevan Nair, and SamanehMadanian . Social Engineering Attacks and Countermeasures. In the New Zealand Banking System: advancing a User-Reflective Mitigation, Information, and Austria. Information 2018,9(5), 110; <https://doi.org/10.3390/info9050110>.
- [12] Aldawood , H.; Skinner, G. Educating and raising awareness on cybersecurity social engineering: Aliterature Review . in proceedings of 2018 IEEE Internation Conference on Teaching, Assessment, and learning for engineering (TALE), Wollongong, NSW, Austria, 4-7 December 2018;pp.62-68.
- [13] Albladi, S.M., Weir, R.S. Predicting individuals' vulnerability to social engineering in social entworks. Cybersecurity, springer open, 2020. <https://doi.org/10.1186/s42400=020-00047-5>.
- [14] Albladi, S.M.;Weir,R.S. User characteristics that influence judgement of social engineering attacks in social networks. Human-centric computing and information sciences (2018). <https://doi.org/10.1186/s13673-018-0128-7>.
- [15] Hussain, Hanizan Shaker; Din, Roshidi; Khidzir, Nik Zulkarnaen; Daud, Khairul Azhar; Ahmad, Suzastri . Risk and Threat via Online Social Network among Academia at Higher Education. International Conference on Big Data and Cloud Computing, 2019.
- [16] Micalleff, N.; Arachchilage,N.A.G. involving users in the design of serious game for security questions education. arXiv preprint. ArXiv: 1710.03888, 2017.
- [17] Soceanu, M. Vasylenko, and A. Gadianru "Improving Cybersecurity Skills Using Network Security Virtual Labs. "In Proceedings of the International Multi Conference of Engineers and Computer Scientists 2017 Vol II, IMECS.
- [18] Merton Lansley, Francois Mouton, Stellios Kapetankis & Nikolaos Polatidis . SEADer ++: social Engineering attack detection in online environments using machine learning. Journal of information and Telecommunication,4:3, 346-362, 2020. doi:10.1080/24751839.202.1747001.
- [19] Nikolaos Tsinganos, Georgios Sakellarios, Panagiotis Fouliras, and IoannisMavridis . Towards an Automated Recognition System for Chatbased Social Engineering Attacks in Enterprise Environments. In ARES 2018: International Conference on Availability, Reliability and Security, August 27-30, 2018, Hamburg, Germany. ACM, ew York, NY USA, 10 pages.
- [20] Mutlaq Alotaibi, Waleed Alfehaid. Information Security Awareness: A Review of Methods, challenges and Solutions.ICITST-WorldCIS-WCST-WCICSS-2018. Information Society. ISBN:978-1-908320-94-0.
- [21] Ana Kovacevic, Sonja, D. Radenkovit. SAWIT: Security Awareness Improvement Tool in Workplace. Applied Sciences, 2020. MDPI. Doi: 10.3390/app10093065.
- [22] Lewin, K. Field Theory in Social Science. Harper and Row: New York, 1951.
- [23] Park, M., Oh,H., Lee, K. Security Risk Measurement for information leakage in IOT-based smart homes from situational awareness perspective. Sensors, 2019,19,2148Met al(2019).
- [24] Abawajy, J. User preferences of cybersecurity awareness delivery methods," Behave. Inf. Technol, vol. 33, no. June 2015, pp-236-247, 2014.
- [25] Ngqoyiyana, IL. Developing an Artefact for Raising S E Awareness among Administrative Staff. A Master of Science in Computer science. Dissertation (2020).

- [26] Wieringa, R. Empirical Research Methods for Technology Validation: Scaling up to practice. *Journal of system and Software* (2013). Doi.org/10.1016/j.js2013.11.1007.
- [27] Vince, Reynolds. *Social Engineering: The art of psychological warfare, human hacking, persuasion and deception* (2015).
- [28] Chapman, A. 2012. Conscious competence learning model: four stages of learning theory- unconscious incompetence, to conscious competence matrix-and other theories and models for learning and change. Businessball, Leiscester, UK. <http://www.businessballs.com/consciouscompetencelearningmodel.htm>.

Evaluating Learning Management System based on PACMAD Usability Model: Brighten Mobile Application

Masyura Ahmad Faudzi, Zaihisma Che Cob, Ridha Omar, Sharul Azim Sharudin

Department of Informatics, College of Computing and Informatics, Universiti Tenaga Nasional, Kajang, Malaysia

Abstract—During the pre-COVID-19 pandemic, mobile learning is just an optional or a supplementary module in learning process. However, when the pandemic hit the world in the middle of 2020, a large number of students were forced to move from traditional learning process to online learning. This has become a critical issue especially for new online learners. Usability of a mobile learning application is important in ensuring that learners are able to learn efficiently and effectively with ease. This study evaluates the usability of the Brighten mobile application; a Moodle-based Learning Management System (LMS) which is currently used by all Universiti Tenaga Nasional's students. The evaluation is based on People at the Center of Mobile Application Development (PACMAD). The results indicate that Brighten mobile application is acceptable in terms of usability's effectiveness, efficiency, learnability, memorability and error-tolerance. Learners' satisfaction level shows a "marginally acceptable" result based on the SUS Adjective Rating Scale and the result for cognitive load shows that the highest cognitive load was in terms of the performance factor.

Keywords—Mobile learning; usability; PACMAD; learning management system; Moodle

I. INTRODUCTION

Mobile learning is the process of learning that allows learners to obtain learning materials anytime and anywhere, using mobile devices, such as mobile phones and tablets. Before the COVID-19 pandemic hits the world in 2020, mobile learning has been used as complimentary learning resources for the traditional in-class teaching [1]. However, the pandemic forced most of the students across the world to rely more on online learning. This can be seen in the increased number of online learning users in Malaysia up from 9.5% to 20.8% in 2020 [2]. In 2021, mobile learning has been seen as an increasingly popular learning method as it is able to improve and make learning easier for students around the world [3].

One of the methods of applying mobile learning technique is through a Learning Management System (LMS). LMS is an application that is used for administering e-learning practices, i.e. planning, implementing and accessing the learning and development programs [4]. One of the most widely used LMS is the Modular Object-Oriented Developmental Learning Environment (Moodle), which is an open-source software that allow personalization of its learning environment [5]. Brighten mobile application is a customized Moodle for students of Universiti Tenaga Nasional (UNITEN) [6]. It is widely used

for online and blended learning in UNITEN. During the peak of COVID-19 pandemic, Brighten has become the main source of learning delivery process in UNITEN.

Usability of a system is when the system can be used by it intended users, in a specified context of use, to achieve goals effectively, efficiently and satisfyingly [7]. Usability testing of applications on mobile device needs to consider challenges such as small screen size [8][9][10], restricted input [11] and design issues [12][13]. People At the Center of Mobile Application Development (PACMAD) Usability Model is a usability model that is developed specifically for measuring mobile application performance based on the seven usability attributes by Harisson, Flood and Duce [14]. It takes into consideration ~~on~~ attributes, which are normally neglected by the other usability models when applied to mobile devices. PACMAD focuses on the effectiveness, efficiency, satisfaction, learnability, memorability, errors and cognitive load factors.

Despite LMS being widely used since 1960s, there are a lot of problems in terms of its usability. These usability problems include inconsistency in design, issues with navigational links and search functions, inappropriate contents and difficult to be use [15]. These issues are very crucial as they might interrupt the effective process of knowledge transfer [16].

The objective of this research is to evaluate the usability of Brighten mobile application using PACMAD usability model. In order to evaluate the usability of the Brighten mobile application, the research questions can be divided into three parts, which are:

RQ1: What is the effectiveness, efficiency, learnability, memorability and errors of the Brighten Mobile Application?

RQ2: What is the satisfaction level of the Brighten Mobile Application's user after using the application?

RQ3: What is the cognitive load of the Brighten Mobile Application's user while using the application?

The remainder of this paper is structured as follows: Section 2 discusses relevant works related to mobile learning and usability testing. This is then followed by Section 3, which explains the details of the methodology used in the experiment. Section 4 discusses the result of the experiment. Finally, in Section 5 provides the conclusions and suggestions for future work.

II. LITERATURE REVIEW

Mobile learning or sometimes known as m-learning is simply defined as a learning process, in terms of pedagogy and education [17], that takes place through mobile devices. Using the mobile technology, mobile learning can be accessed from any location at any time [18]. Ozdamli and Cavus mentioned that this learning method should be ubiquitous, portable, blended, private, interactive, collaborative and instant [19].

Mobile learning is performed through a mobile phone, a tablet, a Personal Digital Assistance (PDA), an iPod, a palmtop or any special ubiquitous handheld devices. Although laptops and notebooks are portable devices, they are not considered as a mobile learning device as they are much bigger and heavier [20].

Mobile learning is applied through mobile medium such as SMS/MMS, email, message boards, forums, blogs and video conferencing. El-Sofany and El-Haggar divided these teaching tools into three categories, which are social networks (such as Facebook, Twitter, Blogs and Youtube EDU), web-based platform (such as Rapid Cycle Evaluation Coach, TED-ed and Moodle) and Internet of Things (such as Smart classroom environment device, attendance system and real-time feedback on lecture quality) [21].

As mobile learning is the extension of e-learning, it inherits all the advantages of e-learning such as supporting distance learning and enhancing student-centered learning [22]. Visual learners gain benefits from mobile learning as compared to learning through textbooks. There are improvements in the communication process between teachers and learners through mobile learning environment. Learners can control their learning process and pace. This leads to efficient learning [23] and positively motivates them to learn [24].

LMS is a system that enables educators to administer, build, track, maintain, update and report information related to a learning program [25][26]. It supports online and offline discussions, formative and summative evaluations and practical-related contents [27]. It assists the learning process in an e-learning environment and can be divided into two (2) types; proprietary LMS and open source LMS [26]. Moodle, Open edX and Chamilo are among the most popular open source LMS [28].

Moodle was developed by Martin Dougiamas and Pete Taylor in 2002 and has established itself as a leading LMS in 2007. Moodle evolved since then and as of 2020, there are more than 190 million of Moodle users around the world [29]. Among the advantage of Moodle compared to the other system is that it can be used by users from different platforms (Windows, Mac, UNIX and Linux) without modification. It supports learning management activity through learning materials, videos, discussion and forums, chat and assessments [26].

Brighten is a customized Moodle that is currently being used in Universiti Tenaga Nasional, Malaysia. It was first implemented in June 2021 to replace the earlier Moodle system. The customization of Moodle into Brighten was done after an informal study conducted on the data usage and data

redundancy of the “older” Moodle LMS. The main interface of the Brighten application is shown in Fig. 1.

Based on International Organization for Standardization (ISO), the evaluation of a system’s effectiveness, efficiency and satisfaction determines the usability of a system [30]. ISO also mentioned that there are three (3) factors that need to be considered when evaluating the usability of a system; user, goal and context of use. Nielsen’s usability model consists of efficiency, satisfaction, learnability, memorability and errors [31].

Despite both of the usability models being widely used, it does not fulfill the context of use when it comes to mobile applications. Hence, PACMAD was introduced to overcome the limitation of the common usability models. PACMAD looks into the context of use for mobile application, such as mobile context, connectivity issues, small screen size, different display resolution, limited processing capability and power and different data entry method [14]. It combines the attributes from both ISO and Nielsen’s usability model and it is designed specifically for mobile application [32]. PACMAD has seven attributes. These attributes are efficiency, effectiveness, learnability, memorability, error, satisfaction and cognitive load. This usability model includes cognitive load as one of the measuring attributes as its main contribution [33]. These attributes are very important for applications that run on mobile devices as mobile devices have different task setting and size limitation compared to desktop PC [34].

Based on these studies, we can conclude that PACMAD is usability model that is created specifically for mobile application. It evaluates the cognitive load of a user. Cognitive load is among the most important factor, which needs to be focused upon when knowledge transfer process takes place. In the current situation where some of the learning process needed to be done online, it is very crucial to have a mobile learning application that can assists in teaching, and not a burden to learners. As such, based on the results of this usability testing, the Brighten mobile application can be further improved.

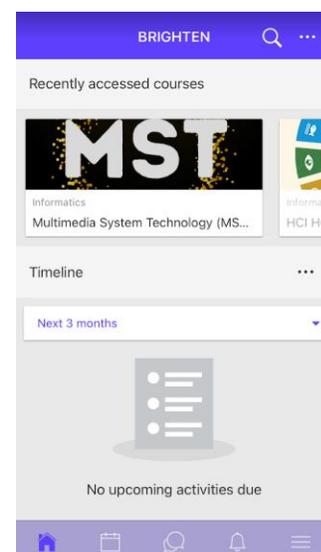


Fig. 1. Brighten Main Page.

III. METHODOLOGY

An experiment was designed which comprises of three (3) tasks and three (3) questionnaires based on the requirements of PACMAD usability model. The participants will evaluate the usability of the Brighten Mobile Application. The experiment was divided into three parts:

- Part 1: Questionnaires to gather participants' information such as age, gender, program of study, year of study, mobile OS, frequency of mobile phone usage per day and average time spends on mobile learning
- Part 2: Evaluate the effectiveness, efficiency, learnability, memorability and error of the tasks that need to be implemented by the participants on Brighten Mobile Application. During this phase, participants are required to perform 3 tasks; Task 1 - sending a private message to the teacher/friend/group, Task 2 - downloading a file (notes, support document, lab manual) and Task 3 - uploading a file (project/lab/assignment submission). For these tasks, participants are required to repeat the process 3 times for different receiver (Task 1) and file types (Task 2 and Task 3). Time will be taken to calculate the efficiency, effectiveness, learnability and memorability attributes. Errors will be calculated manually by the participants during the experiment.
- Part 3: Questionnaires to evaluate the participants' satisfaction level and cognitive load through System Usability Scale (SUS) questionnaire and NASA-TLX, respectively. Participants can add their personal comments about Brighten application and the issues that troubles them during the experiment. The flow of the experiment is as shown in Fig. 2.

The execution of the experiment was done online, through MS-Teams. This was done online, because at that period Malaysia was under the Movement Control Order (MCO) due to COVID-19 pandemic. MCO restricted the mobility of the participants as well as the researchers.

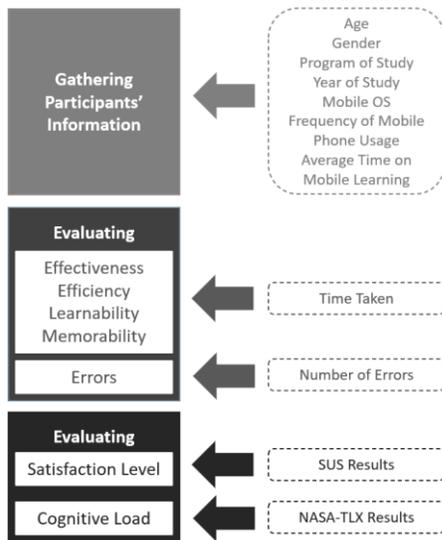


Fig. 2. The Experiment Flow.

IV. RESULT

A. Participants

77 participants were involved in the usability testing. However, 3 of the students pulls out in the middle of the experiment. As such their data are not counted in the result.

All participants are undergraduate students from six different bachelor degree programs (at UNITEN), taking Human Computer Interaction (HCI) subject. Age group, gender, program of study, year of study, mobile OS are the demographics characteristics that have been included in this study as shown in Table I. From 77 participants, more than half of the participants (54%) are between 21 to 23 years old, most of them are male (61%) and almost 80% of the participants are second year students. 61% of the participants are Android users.

Participants were asked on their daily frequency mobile phone usage. A large group of the participants spent between 6 to 12 hours using mobile phone daily as shown in Fig. 3.

Generally, the participants used their mobile phone for communication (WhatsApp, Telegram, Line and Signal), social media (Facebook, Instagram, Twitter and TikTok), mobile learning (Brighten and MS-Teams) and playing games as shown in Fig. 4.

The average students' spending time on mobile learning is shown in Fig. 5. Most participants spent between 1 to 3 hours on mobile phone utilizing mobile learning where they attended lectures, read and download notes, doing quizzes and assignments.

TABLE I. THE DEMOGRAPHICS CHARACTERISTICS OF THE STUDY

Information about the Participants		
Age Group	18-20 years old	30
	21-23 years old	42
	24-26 years old	3
	27 years old and above	2
Gender	Female	30
	Male	47
Program of Study	Bachelor in Computer Science (Hons) (Software Engineering)	25
	Bachelor in Computer Science (Hons) (Cyber Security)	27
	Bachelor in Computer Science (Hons) (System and Networking)	4
	Bachelor in Information Technology (Hons) (Information System)	7
	Bachelor in Information Technology (Hons) (Graphics & Multimedia)	4
	Bachelor in Information Technology (Hons) (Visual Media)	10
Year of study	1	11
	2	61
	3	5
Mobile OS	Android	47
	iOS	30

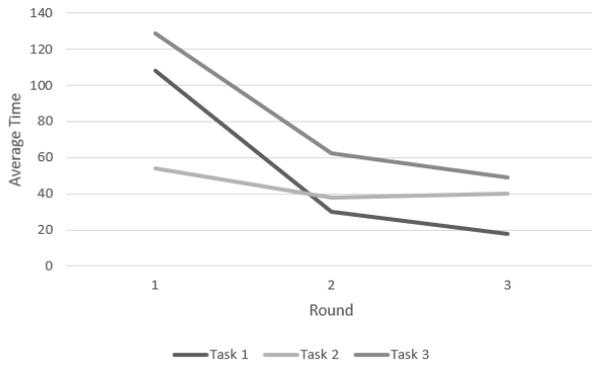


Fig. 3. The Average Time Required for Participants to Complete the Tasks, for each Round.

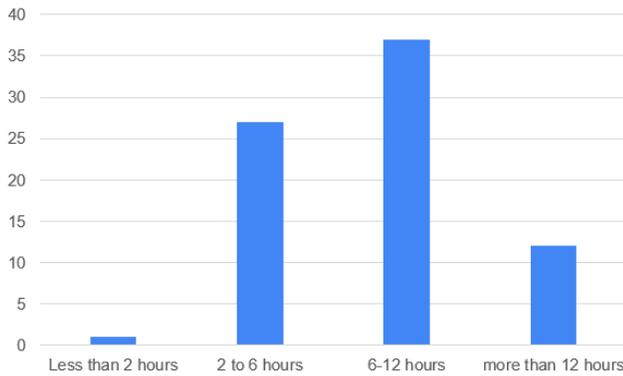


Fig. 4. Frequency of Mobile Phone usage per Day.

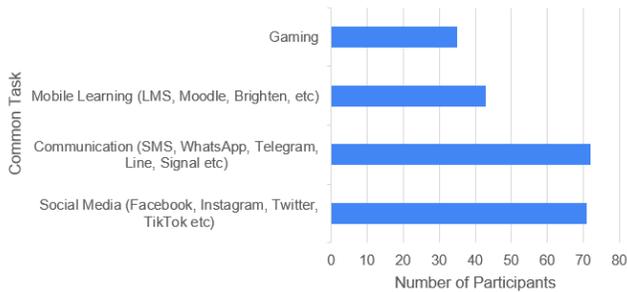


Fig. 5. Common Task on Mobile Phone.

B. Result

The results obtained are based on all the seven elements in PACMAD – effectiveness, efficiency, learnability, memorability, error-tolerance, satisfaction and cognitive load.

1) *Effectiveness*: The result of effectiveness depends on the ability of the participants to complete the tasks given. Table II shows that 98% of the participants managed to complete the tasks in Round 1, 100% in Round 2 and 99% in Round 3. Some of the participants failed in completing the tasks in Round 1 and Round 3 which is either due to network problem or system error that leads to failure in downloading the files.

2) *Efficiency*: Efficiency is calculated based on the speed and accuracy of the participant in completing the tasks measured using the task completion time, as indicated in

Table III. It shows the task per second calculation that demonstrate the number of tasks completed in a second. Overall, the efficiency of the system increases with the number of rounds.

3) *Learnability*: Learnability measures how easy a task is for the users to accomplish it when they first time encounter the interface. It also measures the number of repetitions that the users take to become efficient at that task. Learnability is measured by the time taken for the participants to finish a task and the number of rounds they need in learning on using the system. According to the Nielsen Norman Group, the same task needs to be repeated until the time taken for the participants to finish the task started to plateau. However, in this research, we have fixed the number of rounds to be 3, as mentioned in the Methodology Section.

Several of the participants were eliminated from the calculation as they were facing some technical problem when completing the task. Only 64 results from the participants are included in the calculation for Task 2 and Task 3. The average time required by the participants for each task in each round is shown in Fig. 6.

It can be seen that the time taken for each task is decreasing when it is being repeated for the second and third time.

TABLE II. TASK COMPLETION RESULTS

	Round 1	Round 2	Round 3
Completion Rate	98%	100%	99%

TABLE III. TASK EFFICIENCY

TASK	ROUND		
	1	2	3
1	0.0197 task/second	0.0494 task/second	0.0699 task/second
2	0.0184 task/second	0.0335 task/second	0.0420 task/second
3	0.0202 task/second	0.0515 task/second	0.0545 task/second
Overall relative efficiency	1.94%	4.48%	5.54%

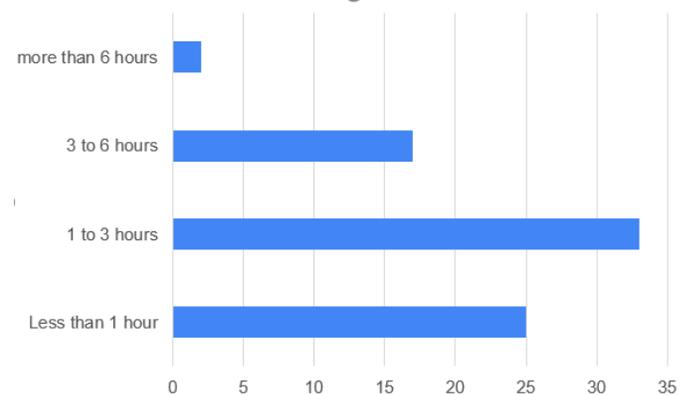


Fig. 6. Average Time Spend on Mobile Learning per Day.

4) *Memorability*: Memorability is about how easy is for the users to reestablish their skills after a long time of no use. Comparing task completion task 1, 2 and 3 on Round 1, Round 2 and Round 3. Based on the task completion on Table II, Round 2 performed the best by 100% had completed the task. Whereas, in Round 1 less than Round 2 which are 98%. Therefore, the task had been increased in Round 3 which are 99% successful completing the task. Mostly student can perform and not give up completing the task. For Round 1, most student spend most time to understand the task given and some students do not understand the task given. But then, they can do the task smoothly for the next round. Issues that affect the performance of the participants are unresponsive application and unstable connection.

5) *Error-tolerance*: Error is measured by calculating the number of errors that are done by the participants in completing the task. Among the error listed are; wrongly entered an input, click the wrong page – need to go to sitemap/menu, click the wrong page – need to press back button, understanding error (doing the wrong task) and press submit button when some questions were left unanswered. Each of the error will be counted including the repetitive errors. If the error faced by the participants is not in the list, participants are required to write the error and how it affects their process in completing the task.

As can be seen in Fig. 7, the number of errors for each of the task can be seem to be decreasing as the number of rounds increasing. The global errors are obtained by dividing the number of errors for a particular round with the number of tasks. Table IV shows the global errors obtained in Round 1, Round 2 and Round 3.

6) *Satisfaction*: SUS questionnaire is used to measure user satisfaction of the apps and the results are shown in Fig. 8. The average score from all respondents is calculated and the results were converted to SUS adjective rating.

The average score from all participants is 61.13. Thus, the result falls in the 'OK' range based on the SUS Adjective Rating Scale.

The participants explained that they are still confused with the "downloading" system during the execution of the task. They are not sure whether they actually have downloaded the file or not. Besides that, the difference between Brighten App via mobile phones and desktop/laptop in terms of UI design also can lead to several confusion and complication. Overall, Brighten App is still acceptable to participant's perceptions of usability value but it needs a bit of improvement in terms of downloading files, IU design, etc.

7) *Cognitive load*: The level of cognitive load involves while using the application is measured to determine the

cognitive processing needs. NASA Task Load Index (TLX) is used to assess work load on five 7-point scales. Increments of high, medium and low instrument consists of five dimensions (one question associated with each dimension) was used to determine the cognitive load of students in performing the given tasks (refer Table V.). The specific dimensions determined the activity's contribution to the cognitive workload and measured using a Likert Scale that range from 1-Very Low and 7-Very High. The overall cognitive workload that the participant experiences is calculated by adding up all the scores and then the average is calculated by dividing the total score with the six different dimensions. The higher score indicated the higher the cognitive workload that the participant experienced.

The highest response is rating 5 with 34% for the performance dimension with the average rating of 5.5, indicating high cognitive load for this dimension. The next average rating of 3.9 for the temporal demand dimension followed by effort, mental, frustration and physical dimension (average scores 3.7, 3.6, 3.4 and 3.2 respectively).

8) *Comments from participants*: Most of the participants managed to complete the tasks without any critical issues. However, some of them has provided some comments and issues on Brighten application that they faced while implementing the tasks. The comments and issues are categorized into five categories. The comments are shown in Fig. 9.

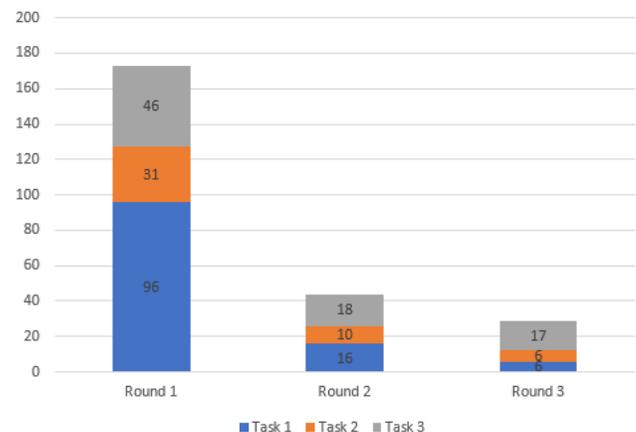


Fig. 7. Total Number of Errors based on Tasks and Rounds.

TABLE IV. GLOBAL ERRORS FOR EACH ROUND

	Round 1	Round 2	Round 3
Total Number of Errors	173	44	29
Global Errors	57.67	14.67	9.67

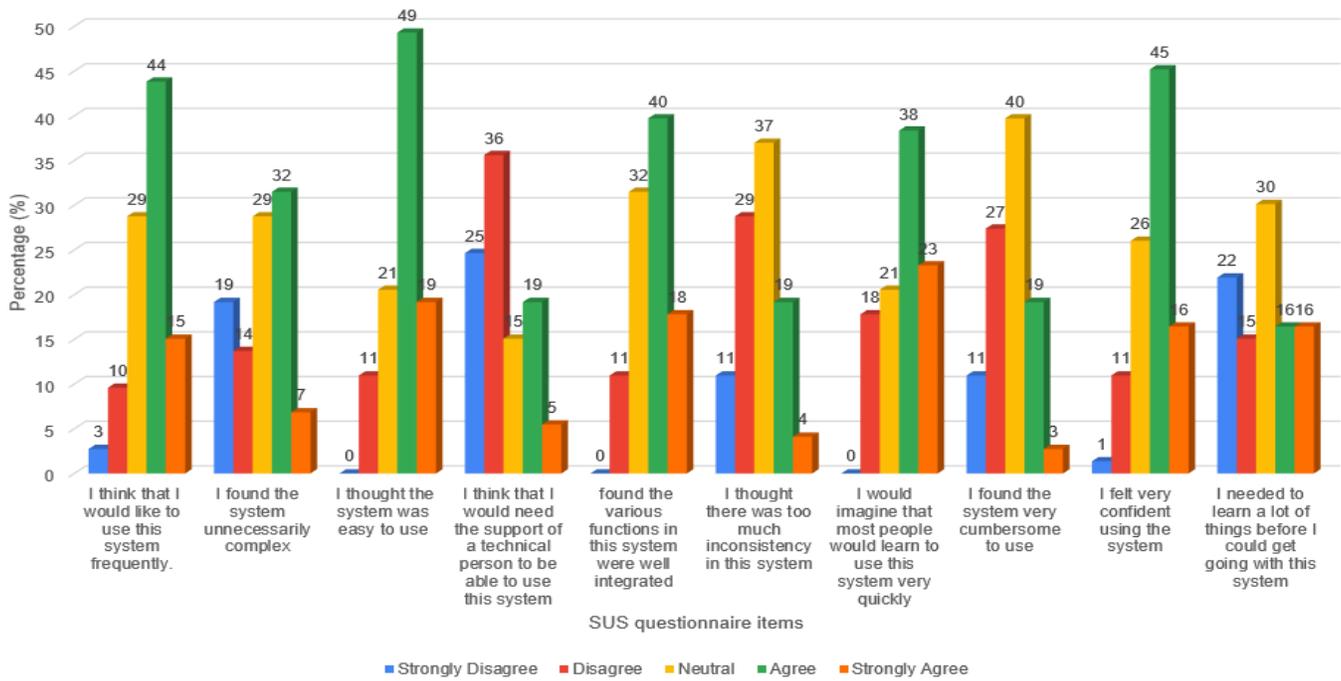


Fig. 8. SUS Questionnaire Score.

TABLE V. THE NASA-TLX RESULT

Dimension	Questions	1	2	3	4	5	6	7
Mental Demand	How mentally demanding was the task?	7%	22%	16%	25%	21%	7%	3%
Physical Demand	How physically demanding was the task?	14%	23%	18%	21%	22%	3%	0
Temporal Demand	How hurried or rushed was the pace of the task?	4%	18%	21%	21%	19%	11%	7%
Performance	How successful were you in accomplishing what you were asked to do?	0	1%	4%	10%	34%	27%	23%
Effort	How hard did you have to work to accomplish your level of performance?	10%	23%	15%	11%	25%	10%	7%
Frustration Level	How insecure, discouraged, irritated, stressed, and annoyed were you?	21%	26%	5%	18%	8%	12%	10%

*The output from these questions will be measured using Likert Scale that range from 1-Very Low and 7-Very High

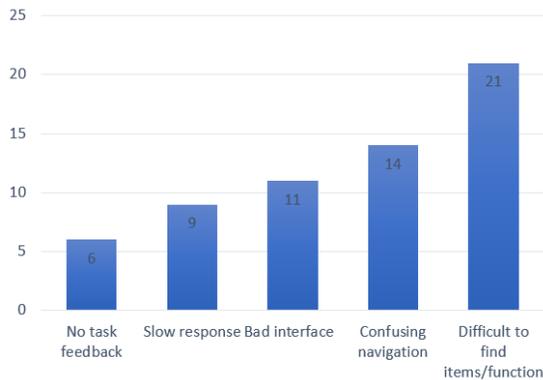


Fig. 9. Comments from Participants.

The most common problem-faced is regarding the difficulties in finding items or functions in the application. In the message sending task, the participants are having problem in finding the send message function or the receiver's name. The application requires the exact full name of the receiver in order to ensure that the message is send to the right person. As the application allows the message to be sent to any user in the university, there is a possibility that the message will be sent to the wrong person.

In the downloading file task, the participants are having problem in finding the location of the files that have been downloaded. The application neither allows the participants to choose the location for the file to be saved nor informing the participants on the location of the saved file.

Some of the participants are having issues in navigating the application. Among the issues that slow down page navigation are; pages kept on reloading and the location of the button is not at common location. Most of these participants find that the mobile version of the application is not as friendly as the web version.

Several participants find that the color template used in the application is too pale and it is difficult to distinguish the sections. This can be seen in Fig. 10.

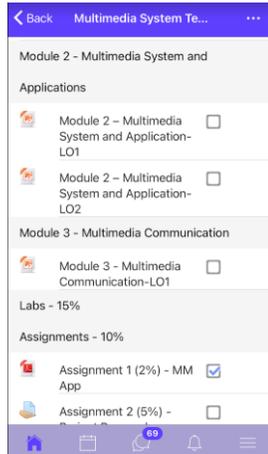


Fig. 10. Brighten Application Subject Page.

Slow response and no task feedback are issues faced by 15 of the participants, where they find that it takes sometimes for the application to respond despite good internet connection. In the task where the participants are required to answer quiz questions, they complained that there is no timer to indicate the time available for them. They find that this is very demotivating.

V. DISCUSSION

From the result, it can be seen that Brighten application passed all of the usability testing with minimum issues. The result of the effectiveness, efficiency, learnability, memorability and errors of the Brighten Mobile Application are acceptable where the learners have no major issues in implementing the tasks.

The average SUS score is 61.13, as mentioned in the previous section. This raw SUS score is below the average of common raw SUS threshold which is 68. This result is deemed as marginally acceptable.

In terms of cognitive load, the participants seemed to have a high cognitive load on performance dimension. This shows that the learners find it quite difficult to successfully accomplishing the given tasks.

In the comment section, some of the participants did mention that they find that the Brighten application on mobile device is difficult to be used compared to the web version. The participants are having problems in finding some of the menu buttons and users.

One of the main issues that interrupted the participants from completing the task is the network issue. Network issues

not only affecting the usability testing result but also the communication between the researcher and the participants.

In this experiment, the number of errors is counted by the participants themselves. To ensure a more reliable result is being obtained, it is suggested that for the experiment to be conducted in face-to-face mode where the researcher can assist the participants in counting the number of errors that they encountered. This can also ensure that the testing is not being interrupted by network issue.

VI. CONCLUSION

This study evaluates the usability of the Brighten application based on PACMAD usability model where it evaluates the application in terms of effectiveness, efficiency, learnability, memorability, error-tolerance, satisfaction and cognitive load. On the surface, the result of the effectiveness, efficiency, learnability, memorability, error, satisfaction of the Brighten mobile application are acceptable with minor issues.

Using SUS and NASA-TLX shows that there are problems in using the application. However, most of the issues are being detailed out in the comment from the participants sections. It was observed that the issues that prevent full satisfaction from the learners and burden the cognitive of the learners comes from the navigation and user interface design.

Future work can be carried out by developing a guideline/framework for mobile learning user interface design that will increase learners' satisfaction and improve their cognitive load.

ACKNOWLEDGEMENT

This study was funded by Yayasan Canselor UNITEN (YCU) Research Grant 2021 (202101014YCU). We would like to thank UNITEN Innovation & Research Management Centre (iRMC) for fund management.

REFERENCES

- [1] L. F. Motiwalla, "Mobile learning: A framework and evaluation," *Comput. Educ.*, vol. 49, no. 3, pp. 581–596, 2007, doi: 10.1016/j.compedu.2005.10.011.
- [2] A. Hani, "Access to mobile phone, computer increased to 98.6%: Stats Dept," 2021. [Online]. Available: <https://themalaysianreserve.com/2021/04/12/access-to-mobile-phone-computer-increased-to-98-6-stats-dept/>. [Accessed: 10-Jun-2021].
- [3] "From Toy To Tool: Cell Phones In Learning - The Tech Edvocate." [Online]. Available: <https://www.thetechedvocate.org/from-toy-to-tool-cell-phones-in-learning/>. [Accessed: 09-Feb-2022].
- [4] R. Sabharwal, M. R. Hossain, R. Chugh, and M. Wells, "Learning Management Systems in the Workplace: A Literature Review," in *IEEE International Conference on Teaching, Assessment, and Learning for Engineering (TALE)*, 2018.
- [5] "About Moodle - MoodleDocs." [Online]. Available: https://docs.moodle.org/311/en/About_Moodle. [Accessed: 23-Nov-2021].
- [6] "UNITEN Learning Management System: Log in to the site." [Online]. Available: <https://brighten.uniten.edu.my/login/index.php>. [Accessed: 23-Nov-2021].
- [7] "ISO 9241-11:2018(en), Ergonomics of human-system interaction — Part 11: Usability: Definitions and concepts." [Online]. Available: <https://www.iso.org/obp/ui/#iso:std:iso:9241:-11:ed-2:v1:en>. [Accessed: 23-Nov-2021].
- [8] T. Alasmari, "The Effect of Screen Size on Students' Cognitive Load in Mobile Learning," *J. Educ. Teaching, Learn.*, vol. 5, no. 2, pp. 280–295, 2020.

- [9] A. Hamzah, A. G. Persada, and A. F. Hidayatullah, "Towards a framework of mobile learning user interface design," *ACM Int. Conf. Proceeding Ser.*, pp. 1–5, 2018, doi: 10.1145/3291078.3291080.
- [10] A. Kaya, R. Ozturk, and C. A. Gumussoy, "Usability Measurement of Mobile Applications with System Usability Scale (SUS)," in *Lecture Notes in Management and Industrial Engineering*, no. January, A. López-Paredes, Ed. Springer, 2019, pp. 389–400.
- [11] S. Hamed and A. Ahmadi, "Survey of designing user interface for mobile applications," *J. Adv. Technol. Eng. Res.*, vol. 3, no. 2, pp. 57–62, 2017, doi: 10.20474/jater-3.2.4.
- [12] J. Kim, J. Kim, Y. Choi, and M. Xia, "Mobile-Friendly Content Design for MOOCs: Challenges, Requirements, and Design Opportunities," in *CHI Conference on Human Factors in Computing Systems*, 2022, pp. 1–6.
- [13] P. Limtrairut, "Newly Developed heuristics to Evaluate M-learning Application Interface," in *2020 - 5th International Conference on Information Technology (InCIT)*, 2020, doi: 10.1109/InCIT50588.2020.9310962.
- [14] R. Harrison, D. Flood, and D. Duce, "Usability of mobile applications: literature review and rationale for a new usability model," *J. Interact. Sci.*, vol. 1, no. 1, p. 1, 2013, doi: 10.1186/2194-0827-1-1.
- [15] L. Hasan, "Usability Problems on Desktop and Mobile Interfaces of the Moodle Learning Management System (LMS)," 2018, doi: 10.1145/3194188.3194192.
- [16] H. Bettayeb, M. T. Alshurideh, and B. Al Kurdi, "The effectiveness of Mobile Learning in UAE Universities: A systematic review of Motivation, Self-efficacy, Usability and Usefulness," *Int. J. Control Autom.*, vol. 13, no. 2, pp. 1558–1579, 2020.
- [17] A. Sattarov and N. Khaitova, "Mobile learning as new forms and methods of increasing the effectiveness of education," *Архив Научных Публикаций Jspi*, vol. 7, no. 12, pp. 1169–1175, 2020.
- [18] B. Biswas, S. K. Roy, and F. Roy, "Students Perception of Mobile Learning during COVID-19 in Bangladesh: University Student Perspective," *Aquademia*, vol. 4, no. 2, p. ep20023, 2020, doi: 10.29333/aquademia/8443.
- [19] F. Ozdamli and N. Cavus, "Basic elements and characteristics of mobile learning," *Procedia - Soc. Behav. Sci.*, vol. 28, pp. 937–942, 2011, doi: 10.1016/j.sbspro.2011.11.173.
- [20] F. Pozzi, "The impact of m-Learning in school contexts: An 'inclusive' perspective," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 4556 LNCS, no. PART 3, pp. 748–755, 2007, doi: 10.1007/978-3-540-73283-9_81.
- [21] H. F. El-Sofany and N. El-Haggar, "The effectiveness of using mobile learning techniques to improve learning outcomes in higher education," *Int. J. Interact. Mob. Technol.*, vol. 14, no. 8, pp. 4–18, 2020, doi: 10.3991/IJIM.V14I08.13125.
- [22] A. Naciri, M. A. Baba, A. Achbani, and A. Kharbach, "Mobile Learning in Higher Education: Unavoidable Alternative during COVID-19," *Aquademia*, vol. 4, no. 1, p. ep20016, 2020, doi: 10.29333/aquademia/8227.
- [23] A. F. Rosmani, A. Abdul Mutalib, and S. M. Zarif, "Hybridising Signaling Principle And Nielsen's Design Guidelines In A Mobile Application," *Asia-Pacific J. Inf. Technol. Multimed.*, vol. 10, no. 02, pp. 62–76, 2021, doi: 10.17576/apjtm-2021-1002-05.
- [24] B. Prasongsap, W. Khaokhajorn, and N. Srisawasdi, "Mobile learning in informal science education: A systematic review from 2010 to 2019," *ICCE 2020 - 28th Int. Conf. Comput. Educ. Proc.*, vol. 2, pp. 425–431, 2020.
- [25] M. A. Alkhateeb and R. A. Abdalla, "Factors influencing student satisfaction towards using learning management system moodle," *Int. J. Inf. Commun. Technol. Educ.*, vol. 17, no. 1, pp. 138–153, 2021, doi: 10.4018/IJICTE.2021010109.
- [26] N. H. S. Simanullang and J. Rajagukguk, "Learning Management System (LMS) Based on Moodle to Improve Students Learning Activity," *J. Phys. Conf. Ser.*, vol. 1462, no. 1, 2020, doi: 10.1088/1742-6596/1462/1/012067.
- [27] G. Gunawan, H. Sahidu, S. Susilawati, A. Harjono, and L. Herayanti, "Learning Management System with Moodle to Enhance Creativity of Candidate Physics Teacher," *J. Phys. Conf. Ser.*, vol. 1417, no. 1, pp. 0–6, 2019, doi: 10.1088/1742-6596/1417/1/012078.
- [28] C. B. Mpungose and S. B. Khoza, "Postgraduate Students' Experiences on the Use of Moodle and Canvas Learning Management System," *Technol. Knowl. Learn.*, no. September, 2020, doi: 10.1007/s10758-020-09475-1.
- [29] "History - MoodleDocs." [Online]. Available: <https://docs.moodle.org/311/en/History>. [Accessed: 25-Nov-2021].
- [30] "ISO - ISO 9241-2:1992 - Ergonomic requirements for office work with visual display terminals (VDTs) — Part 2: Guidance on task requirements." [Online]. Available: <https://www.iso.org/standard/16874.html>. [Accessed: 25-Nov-2021].
- [31] J. Nielsen and M. Kaufmann, "Usability Engineering."
- [32] N. Parsazadeh, R. Ali, M. Rezaei, and S. Z. Tehrani, "The construction and validation of a usability evaluation survey for mobile learning environments," *Stud. Educ. Eval.*, vol. 58, no. August 2017, pp. 97–111, 2018, doi: 10.1016/j.stueduc.2018.06.002.
- [33] W. Ali, O. Riaz, S. Mumtaz, A. R. Khan, T. Saba, and S. A. Bahaj, "Mobile Application Usability Evaluation: A Study Based on Demography," *IEEE Access*, vol. 10, pp. 1–1, 2022, doi: 10.1109/access.2022.3166893.
- [34] A. M. I. Al-amawi, "Usage and Impact of Hotel MobileApplications on Customer Loyalty : The Mediating Role of Customer Satisfaction Usage and Impact of Hotel MobileApplications on Customer Loyalty : The Mediating Role of Customer Satisfaction," no. April, 2022.

Voice Biometrics for Indonesian Language Users using Algorithm of Deep Learning CNN Residual and Hybrid of DWT-MFCC Extraction Features

Haris Isyanto, Ajib Setyo Arifin, Muhammad Suryanegara

Department of Electrical Engineering
Universitas Indonesia, Depok, Indonesia

Abstract—This research develops a Voice Biometrics model for the Indonesian language users by using deep learning algorithm of CNN Residual and Hybrid of DWT-MFCC Feature Extraction. The voice dataset of Indonesian speakers were created with a duration of 5, 10, 15, 20, and 25 minutes. The testing phase of speaker recognition and speech recognition were carried out by comparing the model of CNN Residual with CNN Standard. In the phase of speaker recognition, CNN Residual model has obtained the best results with the highest precision percentage of 99.91% and the highest accuracy of 99.47% at 25 minutes voice samples, compared to the CNN Standard obtaining precision of 96.83% and accuracy of 99.00%. In the phase of speech recognition, CNN Residual model has reached the best performance at 100% accuracy during 20 trials, while CNN Standard only gave 95% accuracy. CNN Residual Model provides a better performance for its accuracy and precision, but it is slightly slower than the CNN Standard, with a time difference of 0.03 – 1.28 seconds.

Keywords—Voice biometric; deep learning; CNN; DWT-MFCC; security

I. INTRODUCTION

The crime of fraud and identity theft has become a crucial threat in cybercrime. It can be associated with the excessive use of the Internet for miscellaneous activities, including online transactions, social networking, and the storage of personal information. To minimize these problems, a biometric identification method was developed, especially for high-level security entry and privacy of sensitive data access in banking transactions [1-3].

The biometric-based personal identification method is one of the alternatives developed especially for high-level security entry, such as government or military buildings, access to sensitive data or information, and theft prevention. Voice biometrics is a biometric technology that utilizes the biological characteristics of the human voice for the identification and authentication of unique patterns for each individual [4-6]. Voice biometrics includes the voice commands, allowing devices such as smartphones, computers, or laptops to receive what the user has spoken and translated into certain electronic commands. The communication of voice commands between the user voice and the device is also known as human-machine interaction [7-9]. The development of the implementation of voice biometrics technology is a solution to maintain the

privacy and security of individual identity data and to avoid frauds.

The voice biometric has been perceived providing a more secure and a more reliable identification and authentication process. In principle, the authentication mechanism can be conducted remotely using a common device such as smartphones and laptops, while the cost of implementing voice biometrics is lower than other biometric solutions because it does not require special devices, such as fingerprint readers or retina scanners. It also has higher security, easy to operate, and accurate identification method to identify a person [10-13].

Currently, voice object recognition research is being developed using the CNN deep learning model. Deep learning Convolutional Neural Network (CNN) technology is one of the neural network algorithms that can assist in solving problems with large amounts of data and data complexity in the object identification process [14-16]. However, most of the previous research provides a separate discussion between the performance of speaker recognition [17-19] and speech recognition [20-22]. Meanwhile, most of the paper discussions on voice biometrics still use machine learning methods, in which it has the disadvantage of not being able to process large amounts of data and not being able to handle the complexity of large data in the identification process of voice biometrics.

In this paper, a Deep Learning voice biometric model was developed using CNN Residual and Hybrid of DWT-MFCC Extraction Feature. The use of CNN Residual is done to simplify the training and validation process, as well as to improve classification accuracy [23, 24]. Meanwhile, the hybrid extraction feature, the Discrete Wavelet Transform (DWT) [25, 26], and Mel-Frequency Cepstral Coefficients (MFCC) extraction features are used to eliminate noise interference, recognize the shape of the voice pattern from a person's characteristics and select the required voices [27, 28].

The test was carried out on 2 security system processes that apply to voice biometrics [29, 30], namely the speaker recognition security system to detect "Whose voice is the person speaking?" [31, 32]. And a speech recognition security system to detect "What keywords are spoken?" [33, 34]. If both securities are successfully accessed, then the system will be "Accepted". But if these two securities fail to be accessed, then the system will be "Rejected". Furthermore, testing is also carried out by measuring the processing time required to carry out a voice biometrics process.

To test the model, this paper compares the performance of the proposed model with the CNN Standard. The comparison is essential to see how the performance of the CNN Residual model may significantly improve the performance of voice biometrics, especially on its accuracy, which lead to better security system.

The remainder of the paper presents Underlying Theories in Section II, elaborates the theory of voice biometric, its relevant studies, and the theory of DWT and MFCC. Section III presents the architecture of the deep learning model, Section IV presents the results and analysis, while Section V concludes this paper.

II. UNDERLYING STUDIES

A. Voice Biometrics

As shown in Fig. 1, the voice biometrics system consists of 2 (two) processes, namely the user enrollment and user verification/authentication [35-38]. The user enrollment is the process of identifying the user's voice identification for registration of the user's voice data into the database. The user enrollment process begins with a capturing process where the user's voice as input is captured by the microphone as a voice sensor. The user's voice input contains the speaker's voice and speech content. Preprocessing is the process of converting analog user input voice signals into digital ones. The process of creating this template is a user identification process that is carried out to register the identity of the user's voice which is a unique individual characteristic, which registers the speaker's voice (speaker recognition) and the speech recognition content which is stored in the database [36]. The workflow of the user enrollment and verification process can be shown in the first line of Fig. 1.

The second process, namely user verification or authentication, is the process of verifying the user's voice by matching the user identification between the incoming voice and the voice that has been registered in the database. In this process, there is a template match process that is intended to verify the voice data by matching the user identification between the incoming voice data and the voice sample data template that has been registered and stored in the previous

database. The output is Voice Biometrics Authentication with validation of the user's voice data (accepted/rejected).

As shown in Fig. 1, in the context of verification, basically the core of the voice biometrics security system works on 2 phases [29, 30], i.e. speaker recognition to detect "Whose voice is the person speaking?" [31, 32], and speech recognition to detect "What keywords are spoken?" [33, 34]. If the voice recognition on both phases are successful, then the system user will be verified, otherwise, it will be rejected. However, most of previous studies discussed the phase of speaker recognition and speech recognition separately.

B. Relevant Studies

The relevant studies of Speaker Recognition, Speech Recognition and their combination on build up the voice biometrics are shown in Table I.

In the previous research, several papers have discussed voice biometric by using machine learning methods, including machine learning k-Nearest Neighbors (k-NN) [35], SVM machine learning, and MFCC extraction features [36], GMM machine learning and MFCC extraction features [37]. However, it is rare for papers to discuss using deep learning algorithms.

The weakness of using deep learning methods ANN and DNN is less reliable than RNN and CNN [48, 49]. While RNN is a sequential data modeling unit, RNN includes less feature compatibility when compared to CNN. The weakness of this RNN has a gradient loss problem. To avoid the problem of disappearing gradients, the RNN is combined with the LSTM. However, this LSTM has the disadvantage that it requires more memory to train [46, 50-52]. CNN is considered more reliable than ANN and RNN. And the weakness of using machine learning methods GMM, GMM-UBM, GMM-HMM, and HMM is only able to process data in smaller amounts and is less able to process complex data [53, 54]. Therefore, the advantages of CNN are having reliable computing capabilities, having a high accuracy, having the ability to process large training data, having an ability to automatically detect important features without human supervision, and being able to classify data complexity in the voice identification process [17-22].

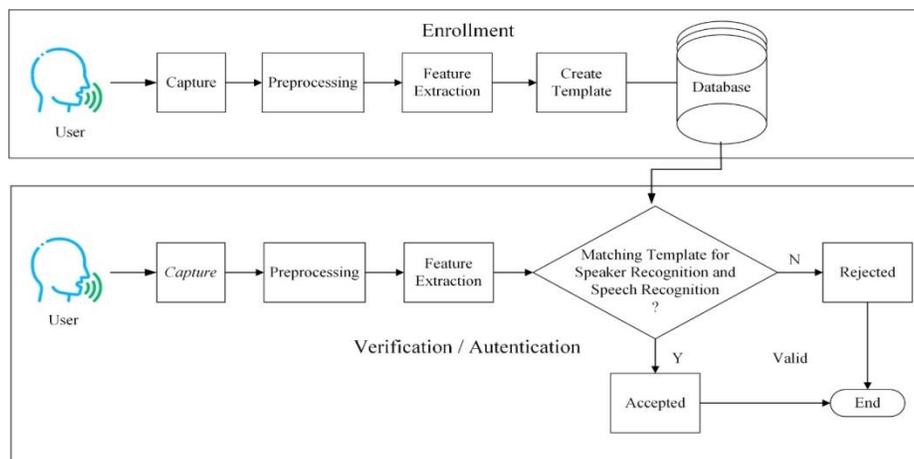


Fig. 1. Process Workflow of user Enrollment and user Verification [36].

TABLE I. RELEVANT STUDIES

Studies of	Research by	Methods	Explanation
Speaker Recognition	[55]	i-vector with machine learning GMM, PNCC, and RASTA PLP feature extraction	PNCC, RASTA PLP and MFCC are sensitive to noise and Machine Learning (ML) is only able to process smaller amounts of data and is less capable of processing complex data
	[56]	i-vector with machine learning GMM-UBM, and MFCC extraction features	
	[57]	i-vektor with deep learning DNN and MFCC extraction features	DNN is less reliable in computing capabilities
	[17-19]	deep Learning CNN	No extraction feature
	[39]	deep-learning CNN and MFCC extraction features	MFCC is sensitive to noise
Speech Recognition	[40]	i-vector with machine learning GMM	ML is only able to process smaller amounts of data and is less capable of processing complex data
	[41]	machine-learning GMM and HMM	
	[42];	machine-learning HMM, MFCC extraction feature	The explanation of ML and MFCC is the same as above
	[43];	machine-learning HMM and deep learning ANN	ANN and DNN are less reliable in computing capabilities
	[44]	deep-learning DNN	
	[45];	deep-learning RNN	RNN only has less feature compatibility compared to CNN
	[46];	deep-learning RNN and LSTM	
	[20-22], [47].	deep-learning CNN and LSTM	No extraction feature
Voice Biometrics	[35];	machine-learning k-Nearest Neighbors (k-NN)	MFCC is sensitive to noise and ML is only able to process smaller amounts of data and is less capable of processing complex data
	[36];	machine learning SVM and MFCC extraction features	
	[37].	machine-learning GMM and MFCC extraction features	
	The Model developed in this research	deep-learning CNN Residual and Hybrid of DWT-MFCC extraction features	Deep-Learning CNN Residual is able to solve problems of large amounts of data, to process complex data, to reduce the number of parameters and arithmetic operations in convolution operations and is able to simplify the training and validation process, thus may increase the classification

			accuracy. In addition, the Hybrid of DWT-MFCC extraction feature can remove noise, recognize the shape of a voice pattern from a person's characteristics, and select the required voices.
--	--	--	--

From the feature extraction point of view, studies in [55, 56] discussed the performance of speaker recognition, while studies in [40-43] discussed the performance of speech recognition with the i-vector extraction features, PNCC, RASTA PLP, and MFCC. They had performed an accuracy of about 76%. Research on speaker recognition [17-19, 39] and speech recognition [20-22, 44-47, 57] by using a deep learning model with i-vector extraction features and MFCC, have obtained an accuracy of around 71-90%.

C. Deep Learning CNN Standard

This CNN standard is a deep learning algorithm technology that has high performance and has been used for database training and testing. With the performance of the CNN standard algorithm, it is expected to improve higher performance compared to using the previous machine learning algorithm.

The architecture of CNN standard is shown in in Fig. 2. It consists of 1 input layer, 5 layers 3x3 Convolution 16 Filters, 1 layer 3x3 Convolution 32/2 Filters, 3 layers 3x3 Convolution 32 Filters, 1 layer 3x3 Convolution 64/2 Filters, 3 layers 3x3 Convolution 64 Filters, 1 layer 3x3 Convolution layer 128/2 Filter, 3 layer 3x3 Convolution 128 Filter, 1 layer Adaptive Average pool, 1 layer Flatten, 1 layer Fully connected and 1 layer output layer.

In CNN Standard, the input layer is a layer for processing data as a set of features. In each Convolution layer, there is a filter/kernel size (filter size) convolution matrix of 3x3 with some filters 16, 32, 64, and 128. This convolutional layer carries the main part of the network computing load, which does most of the heavy computational work. The convolution layer is needed to speed up the extraction of spatial features in the data so that the number of parameters that need to be used to extract features can be reduced, and in the end, will speed up runtime training. Each Convolution layer contains Batch Normalization and ReLu. The batch normalization layer is a normalization technique performed between the layers of the Neural Network, which standardizes the input to the layer for each mini-batch. This is done with mini-batches instead of full datasets. This serves to speed up the training process and uses a higher learning rate, making learning easier. Batch normalization is also able to solve the main problem called internal covariate shift.

ReLU (Rectified Linear Unit) is a node or unit that implements the network layer activation function. ReLU is useful for helping prevent the exponential growth in the computations required to operate neural networks. The adaptive average pooling layer is an easy average pooling operation layer, which gives the input and output dimensions,

to calculate the correct kernel size required to produce an output of the given dimensions from the given input. The flatten layer is a layer that involves taking the combined feature maps generated in the pooling step and converting the data into one-dimensional vectors, to be inserted into the next layer; by flattening the output of the convolution layer to create one long feature vector. And it is connected to the final classification model, which is called the fully-connected layer. Furthermore, a Fully Connected Layer is a layer where all inputs from one layer are connected to each activation unit of the next layer. The last few layers are full connected layers that compile the data extracted by the previous layer to form the final result. The fully connected layer is a full process of Batch Normalization and ReLU data input.

D. MFCC Method for Extraction

As shown in Fig. 1, feature extraction plays an important role to provide good accuracy. Mel Frequency Cepstral Coefficients (MFCC) is believed to be a method that has the highest level of accuracy with speech recognition rates and the fastest feature extraction time compared to other voice feature extraction methods [58]. It is so that the MFCC method is good in accuracy for feature extraction in speech recognition processing in voice biometrics. MFCC is one of the feature extraction methods and methods that are most often used in various fields of voice processing, because it is considered very good in presenting the characteristics of a signal, such as in speech recognition technology, both voice biometrics, speaker recognition, and speech recognition. MFCC is used to recognize the shape of the voice pattern from the extraction of a person's characteristics and choose only the voices that are needed from other voices that are not needed. The feature extraction process with MFCC is a process of taking from feature extraction using a discrete Fourier transform. The Fourier transform can only determine the frequency that appears in a signal, but cannot determine when that frequency appears. The sequence process for the MFCC block diagram can be shown in Fig. 3 [59].

The following is the sequencing process for the MFCC block diagram:

1) *Pre-emphasis*: Used for the filtering process which compensates for the high-frequency portion of the voice signal that is suppressed during the voice production mechanism. The pre-emphasis process is following Equation (1) [28, 59].

$$y(n) = s(n) - a \cdot s(n - 1) \tag{1}$$

where $y(n)$ = signal from the calculation result of pre-emphasis process, n = serial number of voice signal, $s(n)$ = voice signal before pre-emphasis process, a = constant of pre-emphasis filter, with a value of $0.9 \leq a \leq 1.0$ and s = voice signal.

2) *Framing and windowing*: In the framing process, analyze the speech signal of the voice in the form of frames. The signal is divided into several pieces, to facilitate the calculation and analysis of voice signals. Each frame is represented with an interval of 20-40 ms and the signal is continued every 10 ms, which overlaps the previous signal and the next signal [60]. Windowing is used to avoid discontinuity between signals. The most widely used type of window is the hamming window [28, 59, 61].

3) *Fast Fourier Transform (FFT)*: In the Fourier transform, the digital voice signal is transformed into a frequency signal. FFT is an algorithm that has a very fast calculation to perform Fourier transforms in the discrete domain. The results of the FFT process produce detection of frequency domain waves in discrete form [28, 59].

4) *Mel filterbank*: Filterbank is used to determine the energy size of a certain frequency band in a voice signal. Filterbanks are overlapping bandpass filters. Mel is a unit of measure based on the frequency perceived by the human ear. Based on the Mel scale, it is linear below the 1 kHz frequency and logarithmic above it. Mel scaling process according to Equation (2) as follows [27, 59, 60]:

$$mel = 2595 \log_{10} (1 + f / 700) \tag{2}$$

where Mel is the output of the Mel filterbank, and f is the input of the filterbank. While 2595 and 700 are fixed values that have been widely used in the MFCC method in many studies. Mel spaced filterbank as in Fig. 4, the filter bandwidth below 1 kHz is linear while above 10 kHz is logarithmic [59].

5) *Discrete Cosine Transform (DCT)*: DCT is used to calculate the MFCC of a single frame. DCT aims to produce a Mel spectrum to improve recognition quality. The DCT process is following Equation (3) [59].

$$C_m = \sum_{k=1}^K (log_{10} Y[k] \cos [m(k - \frac{1}{2}) \frac{\pi}{K}]); m = 1, 2, \dots \tag{3}$$

In this case, C_m = Coefficient, where $Y[k]$ = the output of the filterbank process on the index, m = the number of coefficients, and K is the expected number of coefficients.

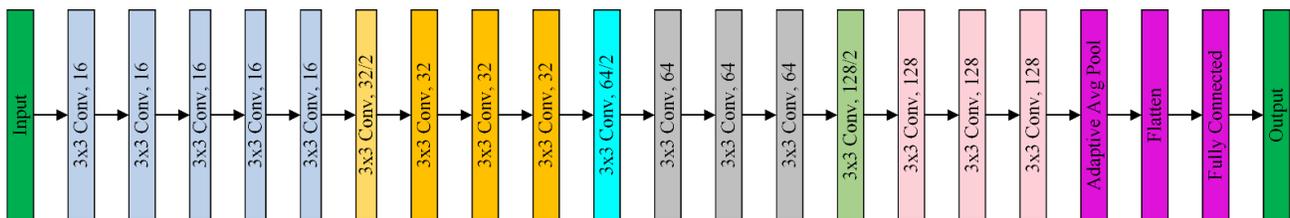


Fig. 2. The Architecture of CNN Standard (CNN No Residual).

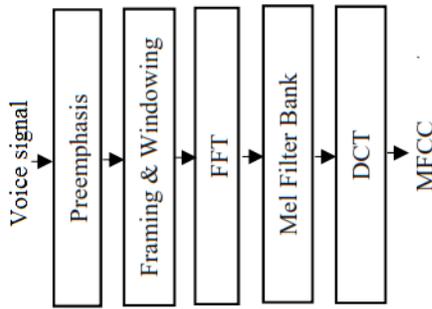


Fig. 3. MFCC Block Diagram [59].

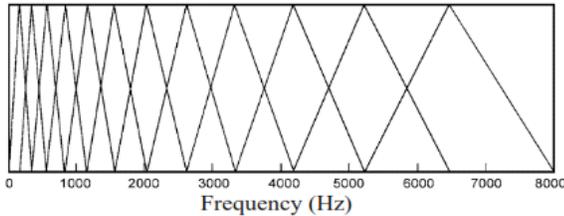


Fig. 4. Example of Mel Spaced Filterbank [59].

6) *Delta coefficients*: Delta coefficients have been used mostly, in addition to the MFCC extraction method. The accuracy of the speech recognition system can be improved by adding the time derivative to obtain stable basic parameters. The equation for calculating the delta can be seen in Equation (4) [59].

$$dt = \frac{\sum_{n=1}^N n(c_{t+n} - c_{t-n})}{2 \sum_{n=1}^N n^2} \quad (4)$$

where dt is the delta coefficients of the t frame. In general, the value of N is 2. The data for the sum of the delta coefficients is the same as the MFCC, the number of coefficients is 13. The sum of the MFCC data plus the delta coefficient is equal to 26 features of the data dimensions [28, 59, 61].

E. DWT Method for Hybrid Extraction MFCC

This MFCC method has drawbacks, where the feature extraction method of voice signals is sensitive to noise [61]. From several previous studies, there is still a need to improve the performance of MFCC. To improve the performance of MFCC on the voice biometrics identification system, a method that can eliminate noise frequencies is needed. There is a need to develop a hybrid method which will help to provide better performance solutions. It is signified that the voice biometrics with a Hybrid DWT-MFCC extraction feature can be used to eliminate noise interference, recognize the shape of the voice pattern from a person's characteristics and choose only the voices that are needed from other voices that are not needed. With the Hybrid MFCC-DWT feature extraction method, it is hoped that reliable features can be formed and produce a high level of accuracy and are better than before [62, 63].

Based on previous research, Discrete Wavelet Transform (DWT) is a good method to eliminate noise (denoising) in signal processing so that the voice quality in voice biometrics is better. The wavelet signal processing is suitable for

nonstationary signals, whose spectral content changes over time. Each wavelet transforms measurement according to a fixed parameter will provide information about the time-temporal range of the signal and information about the frequency spectrum of the signal. The wavelet transform provides an approach to multi-analytical signal resolution and this technique has been used to identify voice signal features. The wavelet transform is an integral part of the raw signal $x(t)$ multiplied by the scale, type shift of the basic wavelet function $\psi(t)$.

Continuous wavelet transform (CWT) is calculated in Equation (5) as follows [26, 63, 64]:

$$CWT(a, b) = \int_R x(t) \frac{1}{\sqrt{a}} \psi^*\left(\frac{t-b}{a}\right) dt \quad (5)$$

where a is the scaling parameter and b is the time localization parameter. DWT is often more efficient than CWT to avoid counting on each CWT scale.

With parameter changes, DWT is defined in Equation (6) as follows [62]:

$$DWT(j, k) = 2^{-j/2} \int_R x(t) \frac{1}{\sqrt{a}} \psi^*(2^{-j} t - k) dt \quad (6)$$

After changing from CWT to DWT, the original continuous wavelet function becomes a discrete wavelet function and a scaling function. The discrete wavelet function and the scaling function as Low Pass Filter (LPF) and High Pass Filter (HPF). The DWT process works like Fig. 5 [63].

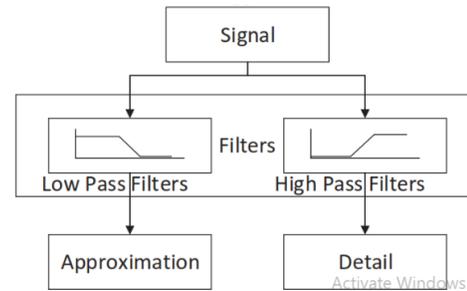


Fig. 5. Process of DWT Signal Filtering [63].

Thus, architecturally it can be described, Referring to Fig. 1, the hybrid extraction process is carried out on the processing results (signal in Fig. 1) which is then carried out with the extra feature DWT-MFCC process to help eliminate noise interference, recognize the shape of the voice pattern from someone and selects the required voices.

III. THE FRAMEWORK

A. The Architecture

Fig. 6 shows the architecture of the voice biometric model which is developed in this research. In principle, the user enrollment/training is processed by using the DWT-MFCC for the part of feature extraction and CNN Residual model for the part of training process. This training process is a capability learning process where the CNN model is trained to identify user voice datasets using large GPU and CPU computing devices. In this training process, the user identification process is carried out to register the user's voice identity which is stored in the database. After completing the training process, the CNN

model that has been trained will produce a Trained CNN Model. Such a trained CNN model will be subsequently used for the user verification process.

This user verification process is the process of classifying and authenticating voice datasets. This user verification process will directly apply the new voice data to the Trained CNN Model and use it to conclude the output. So, when a new user's voice data is entered into the Trained CNN Model, the system will verify the voice data by matching the user identification between the new voice data and the voice data that has been registered in the database. Next, the system will issue predictions based on the prediction accuracy of the data that has been trained on the Trained CNN Model. The Trained CNN Model classification is optimized to maximize prediction performance to achieve high accuracy. The output of the trained CNN model classification is user voice authentication, in the form of data validation (valid / not) or (accepted/rejected) of the user's voice data.

B. CNN Residual as Deep Learning used in this Research

In this research, the architecture of CNN Residual is shown in Fig. 7. The architecture of CNN Residual lies in the implementation of the Residual 8 Shortcut layer by stepping over every 2 layers, which consists of 1 input layer, 5 layers

3x3 Convolution 16 Filters, 1 layer 3x3 Convolution 32/ 2 Filters, 3 layers 3x3 Convolution 32 Filters, 1 layer 3x3 Convolution 64/2 Filters, 3 layers 3x3 Convolution 64 Filters, 1 layer 3x3 Convolution layer 128/2 Filters, 3 layers 3x3 Convolution 128 Filters, 1 layer Adaptive Average pool, 1 Flatten layer, 1 fully connected layer, and 1 output layer.

For CNN Residual, Residual Shortcut is a branching technique for CNN layers, where one branch is a shortcut over 1 or several other branch layers. Initially, the CNN Residual technique was intended to deal with the problem of saturation by increasing the number of layers. Difficult iteration problems and a large number of layers tend to cause a decrease in the quality of classification in terms of speed and accuracy. With the increasing amount of large data, it will affect the increasing capacity of the CNN model, on the number of parameters, filters, and layers. By using the residual technique, the iteration training can be shorter, and the accuracy value will increase, along with the increase in the number of parameters, filters, and layers. The following is the general equation of the shortcut residual identity function, which can be seen in Equations (7) and (8) [24].

$$y = F(x, \{W_i\}) + x \tag{7}$$

$$y = F(x, \{W_i\}) + W_s x \tag{8}$$

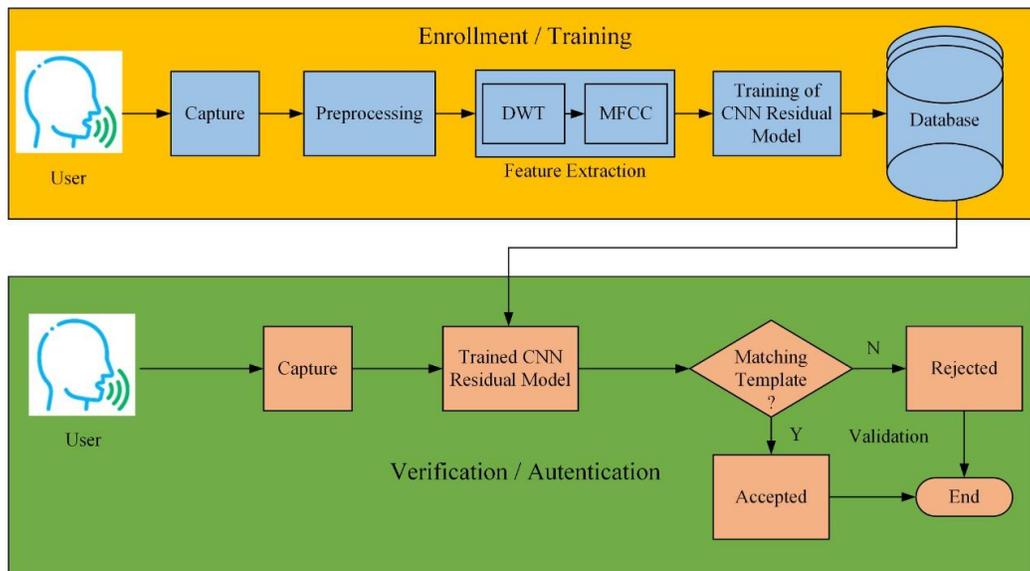


Fig. 6. Framework of Voice Biometrics Model for user Enrollment/Training and user Verification/Authentication Processes, based on DWT-MFCC and CNN Residual.

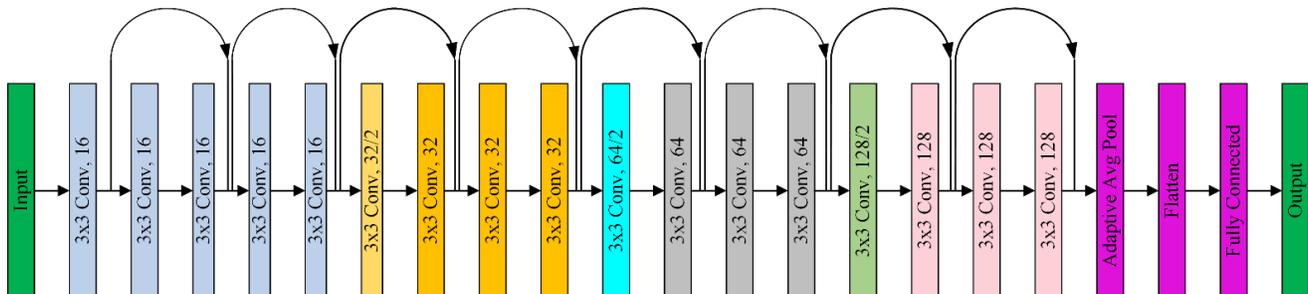


Fig. 7. The Architecture of CNN Residual.

y is a feature map after residual, $F(x, \{W_i\})$ is a filter (residual mapping) whose optimal value is determined, and x is a feature map input. W_i is the layer group that is skipped, and W_s is a linear projection in adjusting the dimensions for x and y when performing shortcuts such as downsampling or upsampling. Although there is almost no change in arithmetic operations and the number of parameters, the addition operation performed can be neglected for the computational load. The application of this residual technique will result in a shorter iteration process and affect the classification results for the better [65, 66]. To further improve the performance of the voice biometrics system, it is proposed to optimize CNN using a CNN residual model. The optimization of this CNN residual is needed to simplify the training and validation process, as well as increase the classification accuracy.

C. CNN Standard as a Comparison of Performance

The performance analysis of CNN Residual model is conducted by comparing with CNN Standard. The essential differences between them are about the Total Parameters and Parameter Size, in which the parameters on CNN Residual Model are greater than the CNN Standard Model. This will affect the working process of the CNN Residual Model which is longer than the CNN Standard Model. CNN Standard Model Parameters and CNN Residual can be seen in Table II.

TABLE II. PARAMETERS OF CNN STANDARD AND CNN RESIDUAL MODEL

No.	Parameter	CNN Standard Model	CNN Residual Model
1.	Total Parameters	707,386	718, 586
2.	Parameter Size (MB)	2.70	2.74

D. Data Set of Indonesian Language

In this research, the original voice dataset of Indonesian language speaker was created. It is essentially used on training the CNN Model algorithm. The creation of the voice dataset begins with the user's voice input via the microphone on the smartphone. The making of this voice dataset involved 10 users, starting from Voice Biometric0 to Voice Biometric9 (VB0 - VB9). Each VB user input contains the unique speaker and speech. Each VB user contributes the voice sample by speaking in Indonesian language for 50 minutes duration.

To make a uniform voice sample files in the dataset, it is necessary to set the following parameters: First, changing the stereo voice to mono voice; Second, changing the frequency of the voice sample rate to 16,000 Hz; Third, truncating the silence to eliminate the pause in the user's voice, so that the result is that every VB user is sampled for 25 minutes, without any pauses; Fourth, segmenting the voice samples for each VB user into 1500 files each; Fifth, changing the voice sample file in the form of a WAV file type format. Finally, with the number of 10 users, a voice dataset is obtained with a total number of voice samples being 15,000 files.

Furthermore, the voice dataset is processed with the DWT-MFCC extraction feature so that it can recognize the shape of the voice pattern from a person's characteristics, can choose only the voices that are needed, and can eliminate noise disturbances. After completing the feature extraction process,

the voice dataset is ready to be trained with the CNN model algorithm.

E. Testing

In this research, the system's performance was tested by conducting a performance assessment.

1) The first phase of Performance Testing, namely Speaker Recognition with the CNN Residual Model Algorithm using DWT-MFCC, (compared to CNN Standard).

2) The second phase of Performance Testing is the performance of Voice Biometric from Speech Recognition with the CNN Residual Model Algorithm using DWT-MFCC, (compared to CNN Standard).

3) Performance Testing of Training Process Time on Voice Biometric with Algorithm CNN Residual Model using DWT-MFCC, (compared to CNN Standard).

Each test was carried out for a sample duration of 5 minutes, 10 minutes, 15 minutes, 20 minutes, and 25 minutes.

IV. RESULT AND DISCUSSION

A. Performance Testing of Speaker Recognition ("Whose Voice is the Person Speaking?")

Performance testing of speaker recognition on the CNN Model is to test the performance of speaker recognition with the CNN Residual Model Algorithm using DWT-MFCC, (compared to CNN Standard). This performance measurement uses the confusion matrix which is a machine learning classification method. This confusion matrix provides information on the comparison of the classification results carried out by the CNN Training model system with the actual classification results. From the results of the CNN Trained Model, it will be used to measure performance with the Confusion Matrix [67, 68]. In this research, a classification system for identifying voice datasets was carried out, where the input data were grouped into 10 VB users to classify the VB voice datasets. In determining the best model, the confusion matrix method becomes important to consider in choosing the best model between deep learning CNN Residual models using DWT-MFCC (compared to CNN Standard).

This performance measurement uses a confusion matrix, which is divided into 4 (four) combinations representing the results of the classification process, namely True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). From the values of TN, FP, FN, and TP, the accuracy and precision of speaker recognition performance with the CNN Standard Model and CNN Residual algorithms are obtained at 5, 10, 15, 20, and 25 minutes of voice sample duration. The analysis data on speaker recognition performance testing with the CNN Standard and CNN Residual Model algorithms can be seen in Table III and IV, and Fig. 8 and 9.

Based on the comparison of accuracy on speaker recognition performance with CNN Residual model using DWT-MFCC, the best results were obtained with the highest percentage accuracy value of 99.47% for the 25 minutes duration voice sample. Accordingly, the larger the number of voice sample durations or the larger the number of voice sample files executed, the higher the percentage of accuracy

performance in prediction. It can be shown in Table III and Fig. 8.

Based on the comparison data of precision on speaker recognition performance with CNN Residual model using DWT-MFCC, the best results were obtained with the highest percentage precision value of 99.91% for the 25 minutes duration voice sample. Based on data analysis, the larger the number of voice sample durations or the larger the number of voice sample files executed, the higher the percentage of precision performance in prediction. It can be shown in Table IV and Fig. 9.

By comparing the performance of Speaker Recognition between CNN Residual Models and CNN Standard, the main results are as follows:

- 1) The accuracy of CNN Residual is higher than CNN Standard, in which CNN Residual is about 96.10% - 99.47% while the later is 95.80% - 99.00%; as can be seen in Table III and Fig. 8.
- 2) The precision of CNN Residual is higher than CNN Standard, in which CNN Residual is about 80.05% - 99.91% while the later is 78.85% - 96.83%; as can be seen in Table IV and Fig. 9.
- 3) The CNN Residual's best results or the highest percentage value are 99.91% precision and 99.47% accuracy for 25 Minutes voice sample duration. The same condition is also applied to the CNN Standard of 96.83% precision and 99.00% accuracy.

It can be implied that the greater the number of voice sample files and the more voice sample training carried out, the higher the level of precision and accuracy in prediction performance will be. By looking at the comparison results, the highest percentage value shows the best value for the precision and accuracy of Speaker Recognition on CNN Residual. It can be concluded that the speaker recognition performance of the CNN Residual model is better than the CNN Standard.

B. Performance Testing of Speech Recognition (“What Keywords are Spoken?”)

Performance testing of speech recognition on the CNN Model Algorithm aims to test the accuracy of speech recognition performance with the CNN Residual Model Algorithm using DWT-MFCC (compared to CNN Standard) at 5, 10, 15, 20, and 25 minutes of voice sample duration. This is done by matching keyword speech or matching speech content.

TABLE III. COMPARISON OF ACCURACY ON SPEAKER RECOGNITION PERFORMANCE WITH CNN STANDARD AND CNN RESIDUAL USING DWT-MFCC

Duration of Voice Samples (Minutes)	Accuracy of Speaker Recognition Performance (%)	
	CNN Standard	CNN Residual
5 Minutes	95,80	96,10
10 Minutes	96,33	96,58
15 Minutes	96,76	97,05
20 Minutes	97,25	97,95
25 Minutes	99,00	99,47

TABLE IV. COMPARISON OF PRECISION ON SPEAKER RECOGNITION PERFORMANCE WITH CNN STANDARD AND CNN RESIDUAL USING DWT-MFCC

Duration of Voice Samples (Minutes)	The precision of Speaker Recognition Performance (%)	
	CNN Standard	CNN Residual
5 Minutes	78,85	80,05
10 Minutes	81,02	82,63
15 Minutes	84,52	86,12
20 Minutes	89,74	93,18
25 Minutes	96,83	99,91

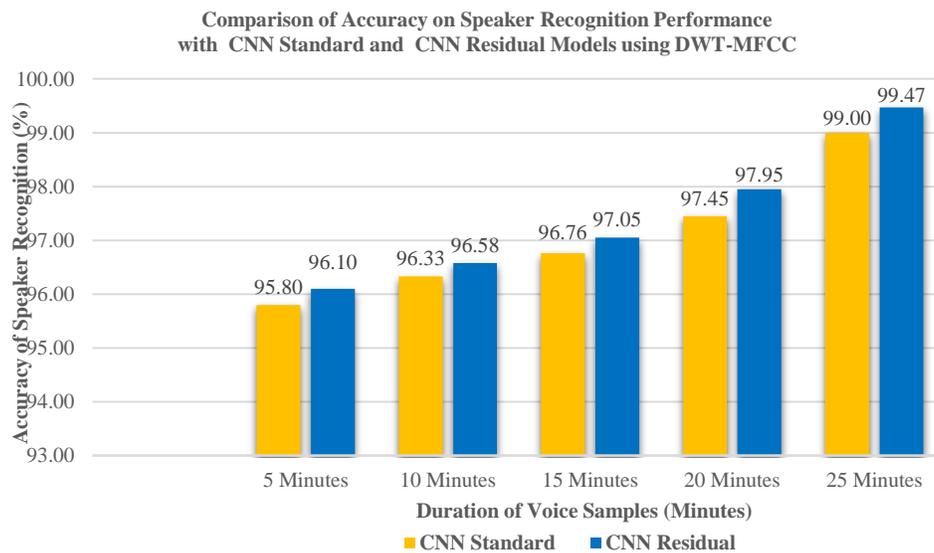


Fig. 8. Comparison of Accuracy on Speaker Recognition Performance with CNN Standard and CNN Residual using DWT-MFCC.

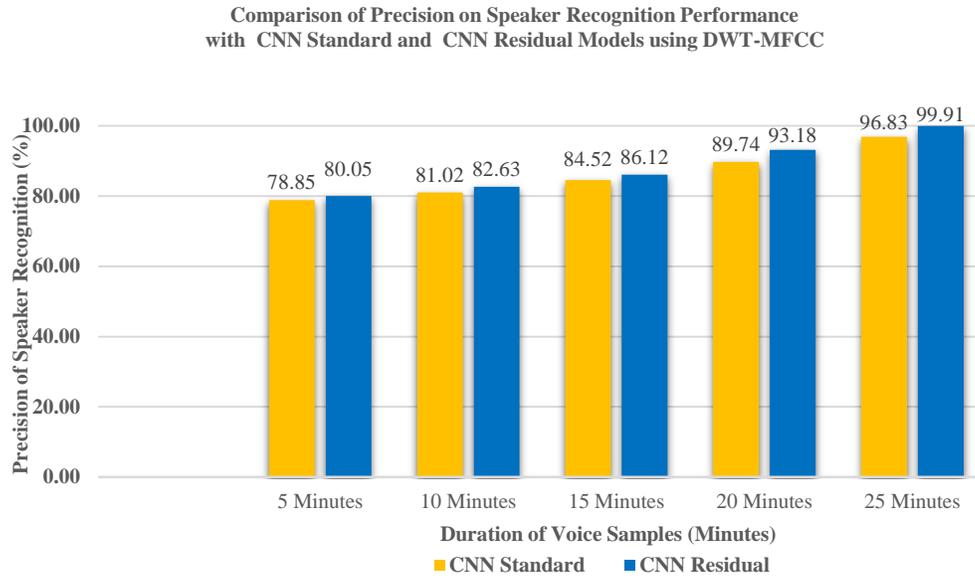


Fig. 9. Comparison of Precision on Speaker Recognition Performance with CNN Standard and CNN Residual using DWT-MFCC.

This test uses a speech content of keyword "Open Access", spoke by the Indonesian users. If the statement is match or correct (True), it will be accepted, while if the speech is wrong or unclear (False), then it is rejected.

Fig. 10 shows that speech recognition performance test has been carried out with the CNN Standard and CNN Residual. It was tested by 20 voice pronunciations, with a total of 10 VB users saying "Open Access". The results show that the percentage of Speech Recognition accuracy performance obtains the best results with the highest percentage value in the

100% CNN Residual, which is higher than the 95% CNN Standard.

The testing has signified that CNN Residual model is better than the CNN Standard. Optimizing the CNN Residual model can improve the validation performance of voice biometric training accuracy, speaker recognition accuracy, and speech recognition accuracy. This is because the CNN Residual model can simplify the training and validation process, as well as increase accuracy in voice biometric classification.

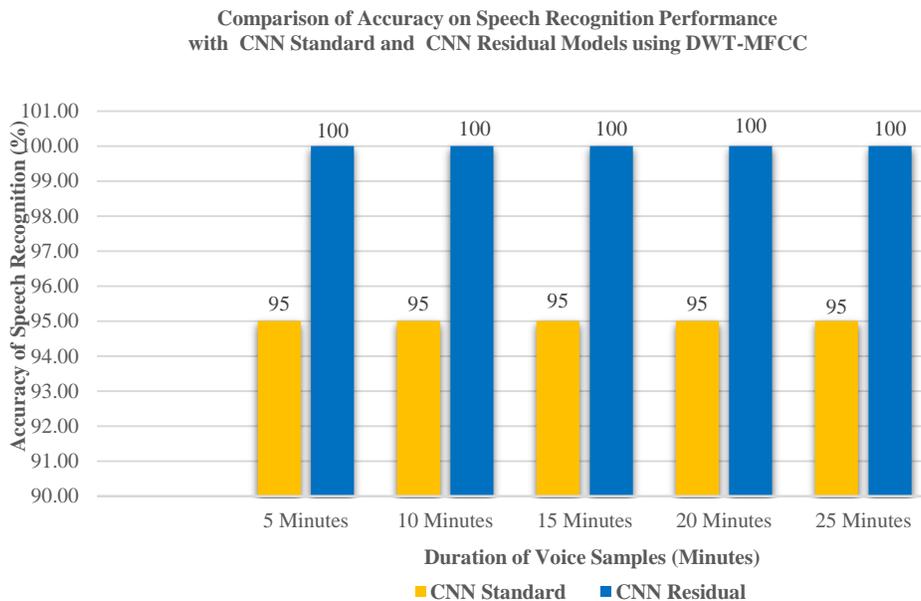


Fig. 10. Comparison of Accuracy of Speech Recognition Performance on Standard CNN Model Algorithm and CNN Residual using DWT-MFCC.

C. Analysis of Time Process for Training

The performance testing of training process time on voice biometrics with CNN Model Algorithm is to test the performance of training process time on the voice biometrics with Algorithm CNN Residual Model using DWT-MFCC, (compared to CNN Standard). This test is to determine how long the processing time is needed for 40 epochs in each running of voice biometrics training on CNN Standard and CNN Residual on voice sample durations of 5, 10, 15, 20 and 25 minutes.

From Fig. 11, the comparison of the performance testing of the voice biometrics training process shows that the CNN Standard training process time performance results are faster than the CNN Residual training process time. This happens because the total number of parameters and the parameter size

of the CNN Residual Model is more than the Standard CNN Model, so it requires a longer processing time, with a time difference of 0.03 – 1.28 seconds. It can be implied that the more training time and the more voice sample files are performed, it will result a higher level of accuracy in prediction. It is also indicated that the larger the file duration, the higher the processing time but with a not too big difference.

By analyzing the performance of training process on voice biometrics for a sample duration of 5, 10, 15, 20, and 25 minutes, it can be signified that the accuracy value are consistently above 95%. Accordingly, it can be concluded that by only using the sample of 5 minutes, the voice biometrics system can recognize and identify the speaker with a decent performance.

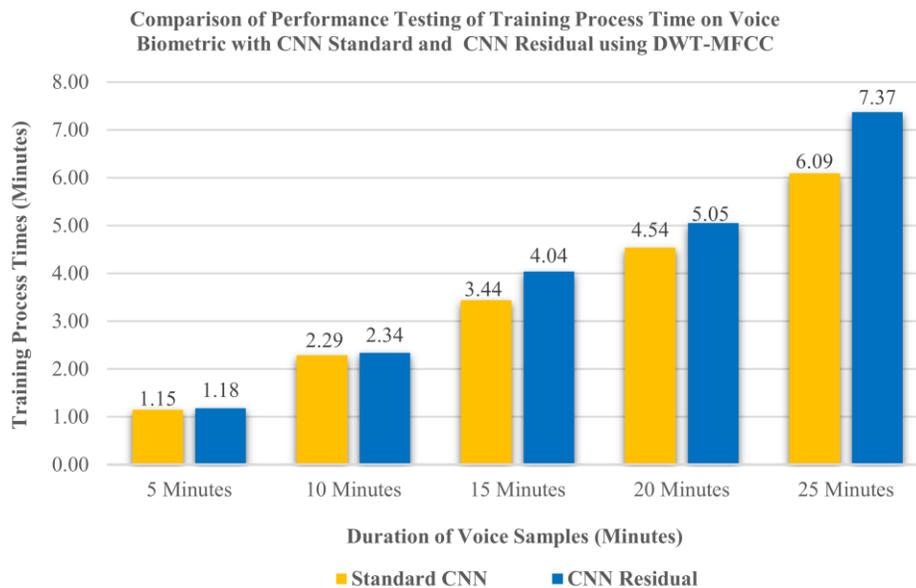


Fig. 11. Comparison of Performance Testing of Training Process Time on Voice Biometric with CNN Standard and CNN Residual Model using DWT-MFCC.

V. CONCLUSION

This paper has developed a Voice Biometric research model for Indonesian language speaker using the CNN Residual Deep Learning algorithm, which uses Hybrid Feature Extraction DWT-MFCC. Testing is done by comparing the model with the CNN Standard. In this study, a voice dataset was created with 10 users (VB0 – VB9). Each VB is a unique speakers who speak in Indonesian language, resulting a total number of 15,000 voice samples with a voice sample duration of 5, 10, 15, 20, and 25 minutes.

The testing was conducted in the phase of speaker recognition and speech recognition. For the speaker recognition phase, the CNN Residual model has obtained the best results with the highest percentage value of 99.91% precision and 99.47% accuracy at a voice sample duration of 25 minutes, compared to Standard CNN of 96.83% precision and 99.00% accuracy. For the speech recognition phase, CNN Residual has achieved the best results of accuracy which is

100% accurate in 20 trials, while Standard CNN only gave 95% accurate results.

From the results of performance testing of training process time for a sample duration of 5, 10, 15, 20, and 25 minutes, the accuracy value has been consistently above 95%. It can be implied that by only using 5 minutes voice data set, this developed system is able to recognize who is the speaker as well as to identify what keywords are spoken.

Optimizing the CNN Residual model can improve the validation performance of voice biometric training accuracy, speaker recognition and speech recognition accuracy as well as its precision. However, CNN Residual is slightly slower than the CNN Standard, with a time difference of 0.03 – 1.28 seconds.

It can be concluded that the performance of the CNN Residual model provides better results for its accuracy and precision. This research is expected to assist in developing a new model that is able to apply an accurate and efficient individual voice identification and authentication algorithm for

voice biometrics systems for security and privacy systems to access sensitive data in banking transactions.

ACKNOWLEDGMENT

Haris Isyanto is in PhD program funded by Beasiswa Pendidikan Pascasarjana Dalam Negeri (BPPDN) Ministry of Education and Culture Republic of Indonesia. Dr. Muhammad Suryanegara is main supervisor, and Dr. Ajib Setyo Arifin is co-supervisor as well as the corresponding author. The voice data set is built by the support of Electrical Engineering - Faculty of Engineering, Universitas Muhammadiyah Jakarta. This publication is supported by Research Grant Universitas Indonesia.

REFERENCES

- [1] Z. Rui and Z. Yan, "A Survey on Biometric Authentication: Toward Secure and Privacy-Preserving Identification," *IEEE Access*, vol. 7, pp. 5994-6009, 2019, doi: 10.1109/ACCESS.2018.2889996.
- [2] S. K. Choudhary and A. K. Naik, "Multimodal Biometric Authentication with Secured Templates — A Review," in 2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI), 23-25 April 2019 2019, pp. 1062-1069, doi: 10.1109/ICOEI.2019.8862563.
- [3] S. Safavi, H. Gan, I. Mporas, and R. Sotudeh, "Fraud Detection in Voice-Based Identity Authentication Applications and Services," in 2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW), 12-15 Dec. 2016 2016, pp. 1074-1081, doi: 10.1109/ICDMW.2016.0155.
- [4] N. A. Kulkarni and L. J. Sankpal, "Efficient Approach Determination for Fake Biometric Detection," in 2017 International Conference on Computing, Communication, Control and Automation (ICCUBEA), 17-18 Aug. 2017 2017, pp. 1-4, doi: 10.1109/ICCUBEA.2017.8463715.
- [5] R. Devi and P. Sujatha, "A study on biometric and multi-modal biometric system modules, applications, techniques and challenges," in 2017 Conference on Emerging Devices and Smart Systems (ICEDSS), 3-4 March 2017 2017, pp. 267-271, doi: 10.1109/ICEDSS.2017.8073691.
- [6] A. Tyagi, Ipsita, R. Simon, and S. K. khatri, "Security Enhancement through IRIS and Biometric Recognition in ATM," in 2019 4th International Conference on Information Systems and Computer Networks (ISCON), 21-22 Nov. 2019 2019, pp. 51-54, doi: 10.1109/ISCON47742.2019.9036156.
- [7] C. Yan, G. Zhang, X. Ji, T. Zhang, T. Zhang, and W. Xu, "The Feasibility of Injecting Inaudible Voice Commands to Voice Assistants," *IEEE Transactions on Dependable and Secure Computing*, vol. 18, no. 3, pp. 1108-1124, 2021, doi: 10.1109/TDSC.2019.2906165.
- [8] S. Sachdev, J. Macwan, C. Patel, and N. Doshi, "Voice-Controlled Autonomous Vehicle Using IoT," *Procedia Computer Science*, vol. 160, pp. 712-717, 2019/01/01/ 2019, doi: <https://doi.org/10.1016/j.procs.2019.11.022>.
- [9] H. Isyanto, A. S. Arifin, and M. Suryanegara, "Design and Implementation of IoT-Based Smart Home Voice Commands for disabled people using Google Assistant," in 2020 International Conference on Smart Technology and Applications (ICoSTA), 20-20 Feb. 2020 2020, pp. 1-6, doi: 10.1109/ICoSTA48221.2020.1570613925.
- [10] A. Boles and P. Rad, "Voice biometrics: Deep learning-based voiceprint authentication system," in 2017 12th System of Systems Engineering Conference (SoSE), 18-21 June 2017 2017, pp. 1-6, doi: 10.1109/SYSOSE.2017.7994971.
- [11] S. Duraibi, F. Sheldon, and W. Alhamdani, "Voice Biometric Identity Authentication Model for IoT Devices," *International Journal of Security, Privacy and Trust Management*, vol. 9, pp. 1-10, 05/31 2020, doi: 10.5121/ijspmt.2020.9201.
- [12] A. Kamalu, A. Raji, and V. I. Nnebedum, "Identity Authentication using Voice Biometrics Technique U," 2015.
- [13] C. Supeshala, "Speaker Recognition using Voice Biometrics," 08/28 2017.
- [14] J. Wang, Y. Zheng, M. Wang, Q. Shen, and J. Huang, "Object-Scale Adaptive Convolutional Neural Networks for High-Spatial Resolution Remote Sensing Image Classification," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 14, pp. 283-299, 2021, doi: 10.1109/JSTARS.2020.3041859.
- [15] P. Fang and Y. Shi, "Small Object Detection Using Context Information Fusion in Faster R-CNN," in 2018 IEEE 4th International Conference on Computer and Communications (ICCC), 7-10 Dec. 2018 2018, pp. 1537-1540, doi: 10.1109/CompComm.2018.8780579.
- [16] X. Wu, R. He, Z. Sun, and T. Tan, "A Light CNN for Deep Face Representation With Noisy Labels," *IEEE Transactions on Information Forensics and Security*, vol. 13, no. 11, pp. 2884-2896, 2018, doi: 10.1109/TIFS.2018.2833032.
- [17] S. Hourri, N. S. Nikolov, and J. Kharroubi, "Convolutional neural network vectors for speaker recognition," *International Journal of Speech Technology*, vol. 24, no. 2, pp. 389-400, 2021/06/01 2021, doi: 10.1007/s10772-021-09795-2.
- [18] H. Salehghaffari, "Speaker Verification using Convolutional Neural Networks," 03/14 2018.
- [19] M. Wang, T. Sirlapu, A. Kwasniewska, M. Szankin, M. Bartscherer, and R. Nicolas, "Speaker Recognition Using Convolutional Neural Network with Minimal Training Data for Smart Home Solutions," in 2018 11th International Conference on Human System Interaction (HSI), 4-6 July 2018 2018, pp. 139-145, doi: 10.1109/HSI.2018.8431363.
- [20] O. Abdel-Hamid, A. Mohamed, H. Jiang, L. Deng, G. Penn, and D. Yu, "Convolutional Neural Networks for Speech Recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 10, pp. 1533-1545, 2014, doi: 10.1109/TASLP.2014.2339736.
- [21] A. Alsobhani, H. Alaboodi, and H. Mahdi, "Speech Recognition using Convolution Deep Neural Networks," *Journal of Physics: Conference Series*, vol. 1973, p. 012166, 08/01 2021, doi: 10.1088/1742-6596/1973/1/012166.
- [22] N. Dimmita and P. Siddaiah, "Speech Recognition Using Convolutional Neural Networks," *International Journal of Engineering and Technology(UAE)*, vol. 7, pp. 133-137, 09/25 2018, doi: 10.14419/ijet.v7i4.6.20449.
- [23] S. T. Seydi, M. Hasanlou, M. Amani, and W. Huang, "Oil Spill Detection Based on Multiscale Multidimensional Residual CNN for Optical Remote Sensing Imagery," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 14, pp. 10941-10952, 2021, doi: 10.1109/JSTARS.2021.3123163.
- [24] E. Ihsanto, K. Ramli, D. Sudiana, and T. S. Gunawan, "Fast and Accurate Algorithm for ECG Authentication Using Residual Depthwise Separable Convolutional Neural Networks," *Applied Sciences*, vol. 10, no. 9, p. 3304, 2020. [Online]. Available: <https://www.mdpi.com/2076-3417/10/9/3304>.
- [25] Y. Cengiz and Y. D. U. Aröz, "An Application for speech denoising using Discrete wavelet transform," in 2016 20th National Biomedical Engineering Meeting (BIYOMUT), 3-5 Nov. 2016 2016, pp. 1-4, doi: 10.1109/BIYOMUT.2016.7849377.
- [26] M. F. Pouyani, M. Vali, and M. A. Ghasemi, "Lung sound signal denoising using discrete wavelet transform and artificial neural network," *Biomedical Signal Processing and Control*, vol. 72, p. 103329, 2022/02/01/ 2022, doi: <https://doi.org/10.1016/j.bspc.2021.103329>.
- [27] S.-Y. Jung, C.-H. Liao, Y.-S. Wu, S.-M. Yuan, and C.-T. Sun, "Efficiently Classifying Lung Sounds through Depthwise Separable CNN Models with Fused STFT and MFCC Features," *Diagnostics*, vol. 11, no. 4, p. 732, 2021. [Online]. Available: <https://www.mdpi.com/2075-4418/11/4/732>.
- [28] A. Antony and R. Gopikakumari, "Speaker identification based on combination of MFCC and UMRT based features," *Procedia Computer Science*, vol. 143, pp. 250-257, 2018/01/01/ 2018, doi: <https://doi.org/10.1016/j.procs.2018.10.393>.
- [29] K. Khotimah et al., "Validation of Voice Recognition in Various Google Voice Languages using Voice Recognition Module V3 Based on Microcontroller," in 2020 Third International Conference on Vocational Education and Electrical Engineering (ICVEE), 3-4 Oct. 2020 2020, pp. 1-6, doi: 10.1109/ICVEE50212.2020.9243184.

- [30] S. Fegade, D. Chaturvedi, and D. Agarwal, "Voice Recognition Technology : A Review," *International Journal of Advanced Research in Science, Communication and Technology*, pp. 31-34, 08/03 2021, doi: 10.48175/IJARSCT-1807.
- [31] Z. Bai and X.-L. Zhang, "Speaker recognition based on deep learning: An overview," *Neural Networks*, vol. 140, pp. 65-99, 2021/08/01/ 2021, doi: <https://doi.org/10.1016/j.neunet.2021.03.004>.
- [32] I. Bisio, C. Garibotto, A. Grattarola, F. Lavagetto, and A. Sciarone, "Smart and Robust Speaker Recognition for Context-Aware In-Vehicle Applications," *IEEE Transactions on Vehicular Technology*, vol. 67, no. 9, pp. 8808-8821, 2018, doi: 10.1109/TVT.2018.2849577.
- [33] Đ. T. Grozdić and S. T. Jovičić, "Whispered Speech Recognition Using Deep Denoising Autoencoder and Inverse Filtering," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 12, pp. 2313-2322, 2017, doi: 10.1109/TASLP.2017.2738559.
- [34] H. Isyanto, A. S. Arifin, and M. Suryanegara, "Performance of Smart Personal Assistant Applications Based on Speech Recognition Technology using IoT-based Voice Commands," in *2020 International Conference on Information and Communication Technology Convergence (ICTC)*, 21-23 Oct. 2020 2020, pp. 640-645, doi: 10.1109/ICTC49870.2020.9289160.
- [35] L. Moreno, "The Voice Biometrics Based on Pitch Replication," *International Journal for Innovation Education and Research*, vol. 6, pp. 351-358, 10/31 2018, doi: 10.31686/ijer.Vol6.Iss10.1201.
- [36] S. Duraibi, F. T. Sheldon, and W. Alhamdani, "Voice Biometric Identity Authentication Model for IoT Devices," *International Journal of Security, Privacy and Trust Management*, vol. 9, pp. 1-10, 05/31 2020, doi: 10.5121/ijspmt.2020.9201.
- [37] S. al, "Voice Biometric: A Novel and Realistic Approach," *Turkish Journal of Computer and Mathematics Education (TURCOMAT)*, vol. 12, pp. 5684-5694, 04/10 2021, doi: 10.17762/turcomat.v12i3.2243.
- [38] X. Zhang, Q. Xiong, Y. Dai, and X. Xu, "Voice Biometric Identity Authentication System Based on Android Smart Phone," in *2018 IEEE 4th International Conference on Computer and Communications (ICCC)*, 7-10 Dec. 2018 2018, pp. 1440-1444, doi: 10.1109/CompComm.2018.8780990.
- [39] A. Chowdhury and A. Ross, "Fusing MFCC and LPC Features Using 1D Triplet CNN for Speaker Recognition in Severely Degraded Audio Signals," *IEEE Transactions on Information Forensics and Security*, vol. 15, pp. 1616-1629, 2020, doi: 10.1109/TIFS.2019.2941773.
- [40] J. Gomes and M. El-Sharkawy, *i-Vector Algorithm with Gaussian Mixture Model for Efficient Speech Emotion Recognition*. 2015, pp. 476-480.
- [41] D. Hoesen, C. Satriawan, D. Lestari, and M. L. Khodra, "Towards Robust Indonesian Speech Recognition with Spontaneous-Speech Adapted Acoustic Models," *Procedia Computer Science*, vol. 81, pp. 167-173, 12/31 2016, doi: 10.1016/j.procs.2016.04.045.
- [42] S. Gholamdokht-Firooz, F. Almasganj, and Y. Shekofteh, "Improvement of automatic speech recognition systems via nonlinear dynamical features evaluated from the recurrence plot of speech signals," *Computers & Electrical Engineering*, vol. 58, pp. 215-226, 02/01 2017, doi: 10.1016/j.compeleceng.2016.07.006.
- [43] Q. Liu, Z. Chen, H. Li, M. Huang, Y. Lu, and K. Yu, "Modular End-to-End Automatic Speech Recognition Framework for Acoustic-to-Word Model," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. PP, pp. 1-1, 07/15 2020, doi: 10.1109/TASLP.2020.3009477.
- [44] Y.-f. Liao and Y.-R. Wang, *Some Experiences on Applying Deep Learning to Speech Signal and Natural Language Processing*. 2018, pp. 83-94.
- [45] M. Elmahdy and A. Morsy, *Subvocal Speech Recognition via Close-Talk Microphone and Surface Electromyogram Using Deep Learning*. 2017, pp. 165-168.
- [46] Z. Ma, Y. Liu, X. Liu, J. Ma, and F. Li, "Privacy-Preserving Outsourced Speech Recognition for Smart IoT Devices," *IEEE Internet of Things Journal*, vol. 6, no. 5, pp. 8406-8420, 2019, doi: 10.1109/JIOT.2019.2917933.
- [47] W. Xiong, L. Wu, F. Alleva, J. Droppo, X. Huang, and A. Stolcke, "The Microsoft 2017 Conversational Speech Recognition System," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 15-20 April 2018 2018, pp. 5934-5938, doi: 10.1109/ICASSP.2018.8461870.
- [48] N. T. Babu, A. Aravind, A. Rakesh, M. Jahzan, R. D. Prabha, and M. Ramalinga Viswanathan, "Automatic fault classification for journal bearings using ANN and DNN," *Archives of Acoustics*, vol. 43, pp. 727-738, 01/01 2018, doi: 10.24425/aoa.2018.125166.
- [49] O. I. Abiodun et al., "Comprehensive Review of Artificial Neural Network Applications to Pattern Recognition," *IEEE Access*, vol. 7, pp. 158820-158846, 2019, doi: 10.1109/ACCESS.2019.2945545.
- [50] S. A. Syed, M. Rashid, S. Hussain, and H. Zahid, "Comparative Analysis of CNN and RNN for Voice Pathology Detection," *BioMed Research International*, vol. 2021, p. 6635964, 2021/04/15 2021, doi: 10.1155/2021/6635964.
- [51] I. Banerjee et al., "Comparative effectiveness of convolutional neural network (CNN) and recurrent neural network (RNN) architectures for radiology text report classification," *Artificial Intelligence in Medicine*, vol. 97, pp. 79-88, 2019/06/01/ 2019, doi: <https://doi.org/10.1016/j.artmed.2018.11.004>.
- [52] C. Zhao, J. Han, and X. Xu, "CNN and RNN Based Neural Networks for Action Recognition," *Journal of Physics: Conference Series*, vol. 1087, p. 062013, 09/01 2018, doi: 10.1088/1742-6596/1087/6/062013.
- [53] D. Chauhan et al., "Comparison of machine learning and deep learning for view identification from cardiac magnetic resonance images," *Clinical Imaging*, vol. 82, pp. 121-126, 2022/02/01/ 2022, doi: <https://doi.org/10.1016/j.clinimag.2021.11.013>.
- [54] C. Janiesch, P. Zschech, and K. Heinrich, "Machine learning and deep learning," *Electronic Markets*, vol. 31, no. 3, pp. 685-695, 2021/09/01 2021, doi: 10.1007/s12525-021-00475-2.
- [55] P. K. Nayana, D. Mathew, and A. Thomas, "Comparison of Text Independent Speaker Identification Systems using GMM and i-Vector Methods," *Procedia Computer Science*, vol. 115, pp. 47-54, 12/31 2017, doi: 10.1016/j.procs.2017.09.075.
- [56] A. Poddar, M. Sahidullah, and G. Saha, "Speaker Verification with Short Utterances: A Review of Challenges, Trends and Opportunities," *IET Biometrics*, vol. 7, 10/03 2017, doi: 10.1049/iet-bmt.2017.0065.
- [57] J. Zhong, W. Hu, F. Soong, and H. Meng, *DNN i-Vector Speaker Verification with Short, Text-Constrained Test Utterances*. 2017, pp. 1507-1511.
- [58] S. Tantisatirapong, C. Prasoproeck, and M. Phothisonothai, "Comparison of Feature Extraction for Accent Dependent Thai Speech Recognition System," in *2018 IEEE Seventh International Conference on Communications and Electronics (ICCE)*, 18-20 July 2018 2018, pp. 322-325, doi: 10.1109/CCE.2018.8465705.
- [59] A. Winursito, R. Hidayat, and A. Bejo, "Improvement of MFCC feature extraction accuracy using PCA in Indonesian speech recognition," in *2018 International Conference on Information and Communications Technology (ICOIACT)*, 6-7 March 2018 2018, pp. 379-383, doi: 10.1109/ICOIACT.2018.8350748.
- [60] N. Chauhan, T. Isshiki, and D. Li, "Speaker Recognition Using LPC, MFCC, ZCR Features with ANN and SVM Classifier for Large Input Database," in *2019 IEEE 4th International Conference on Computer and Communication Systems (ICCCS)*, 23-25 Feb. 2019 2019, pp. 130-133, doi: 10.1109/CCOMS.2019.8821751.
- [61] U. Bhattacharjee, S. Gogoi, and R. Sharma, "A statistical analysis on the impact of noise on MFCC features for speech recognition," in *2016 International Conference on Recent Advances and Innovations in Engineering (ICRAIE)*, 23-25 Dec. 2016 2016, pp. 1-5, doi: 10.1109/ICRAIE.2016.7939548.
- [62] A. K. H. Al-Ali, D. Dean, B. Senadji, V. Chandran, and G. R. Naik, "Enhanced Forensic Speaker Verification Using a Combination of DWT and MFCC Feature Warping in the Presence of Noise and Reverberation Conditions," *IEEE Access*, vol. 5, pp. 15400-15413, 2017, doi: 10.1109/ACCESS.2017.2728801.
- [63] M. Abdalla, H. Abobakr, and T. Gaafar, "DWT and MFCCs based Feature Extraction Methods for Isolated Word Recognition," *International Journal of Computer Applications*, vol. 69, pp. 21-25, 05/17 2013, doi: 10.5120/12087-8165.

- [64] N. Mukherjee, A. Chattopadhyaya, S. Chattopadhyay, and S. Sengupta, "Discrete-Wavelet-Transform and Stockwell-Transform-Based Statistical Parameters Estimation for Fault Analysis in Grid-Connected Wind Power System," *IEEE Systems Journal*, vol. 14, no. 3, pp. 4320-4328, 2020, doi: 10.1109/JSYST.2020.2984132.
- [65] M. Ogawa and Y. Yang, "Residual-Network -Based Deep Learning for Parkinson's Disease Classification using Vocal Datasets," in *2021 IEEE 3rd Global Conference on Life Sciences and Technologies (LifeTech)*, 9-11 March 2021 2021, pp. 275-277, doi: 10.1109/LifeTech52111.2021.9391925.
- [66] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 27-30 June 2016 2016, pp. 770-778, doi: 10.1109/CVPR.2016.90.
- [67] P. Shih, J. Wang, H. Lee, H. Kai, H. Kao, and Y. Lin, "Notice of Violation of IEEE Publication Principles: Acoustic and Phoneme Modeling Based on Confusion Matrix for Ubiquitous Mixed-Language Speech Recognition," in *2008 IEEE International Conference on Sensor Networks, Ubiquitous, and Trustworthy Computing (sutc 2008)*, 11-13 June 2008 2008, pp. 500-506, doi: 10.1109/SUTC.2008.78.
- [68] J. Xu, Y. Zhang, and D. Miao, "Three-way confusion matrix for classification: A measure driven view," *Information Sciences*, vol. 507, pp. 772-794, 2020/01/01/ 2020, doi: <https://doi.org/10.1016/j.ins.2019.06.064>.

Hybrid Deformable Convolutional with Recurrent Neural Network for Optimal Traffic Congestion Prediction: A Fuzzy Logic Congestion Index System

Sara Berrouk, Abdelaziz El Fazziki, Mohammed Sadgal
Computer Science Engineering Laboratory
Faculty of Sciences, Cadi Ayyad University, Marrakech, Morocco

Abstract—In the field of Intelligent Transportation Systems (ITS), traffic congestion is considered as an important problem. Traffic blockage usually affects the quality of time, travel time, economy of the country, and transportability of people. The information of traffic congestion is collected and analyzed in ITS, and the methods to prevent the traffic congestion are predicted. However, the tackling of huge data is still challenging. The rapid increase in vehicle usage and road construction has resulted in traffic congestion. Various studies are undergone in ITS to recognize the traffic management system by adopting few resources. Real time-based traffic services are implemented to prevent the traffic congestion in existing areas. These services provide high expense accuracy. This paper plans to develop a new technique to predict the traffic congestion using improved deep learning approaches. At first, the benchmark dataset is gathered and the pre-processing of data is performed with removing the bad data, organizing the raw data, and filling the null values. The optimized weighted features are selected from the pre-processed data by adopting a new meta-heuristic Hybrid Jaya Harris Hawk Optimization (HJHHO) algorithm. The prediction of congestion parameters such as speed reduction rate, very low speed rate, and volume to capacity ratio of vehicles are performed by the proposed Improved Deformable Convolutional Recurrent Network (IDCRN) prediction model. These predicted measures are subjected to fuzzy interference system for congestion index computation. From the experimental analysis, it has proved that the proposed method has reduced the error rate while comparing with other deep learning and machine learning approaches.

Keywords—Optimal traffic congestion prediction; deformable convolutional network; recurrent neural network; Hybrid Jaya Harris Hawk Optimization Algorithm; congestion index computation; fuzzy interference system

I. INTRODUCTION

Population growth and the rapid urbanization in economy have increased the traffic clogging drastically in most developing cities and large areas in all over the world. It also increases both the road rage statistics and commute time. Consequently, the cause of accidents is increasing due to high traffic congestion [1]. Nowadays, the study for traffic management is highly significant among the researchers. The traffic clogging can be prevented by developing the infrastructure of transportation, which is expensive or else by organizing possible traffic schemes, like analyzing the blocking pattern or prediction of traffic congestion in short

term that can be more effective for road networks in a short time [2].

The timely congestion prediction models focus on the traffic factors such as speed, volume, and traffic stream on group of roads, one way road and especially small rural roads due to the unavailability of data. The limitation on predicting the road networks causes trouble for both the traffic agencies and commuters [3]. The researchers use the data, which is gathered from sensors like CCTV cameras, road sensors etc. These road sensors are located everywhere in road and even in vehicle networks to operate in all routes. This type of data is inconvenient because the process of operating, installation and maintenance is difficult, and it is expensive as it requires permission to access the data for third parties [4]. The real time traffic information like the average speed and blocking level on the section of roadway are provided in the web services such as Seoul Transportation Operation, Google Traffic, Information Service (TOPIS), Baidu Map and Bing Map [5]. Therefore, these web services are freely accessible and offer traffic congestion-related information to the majority cities all over the world, but only few studies are available based on this web services. The data requires multiple inputs, which are in time series, the processing of numerous input traffic images is complex due to high cost [6].

In past years, the data-driven methods were followed by the researchers to develop the mathematical and statistical models to detect the relation of time series in traffic data, also known as parametric method [7]. Mainly, the existing works were related to the presumption of stationary and linearity to capture the prediction trends such as smoothing model, error component method, and historical average method [8]. The seasonal pattern and long-term trends are decomposed to detect the pattern using Autoregressive Integrated Moving Average (ARIMA) model. However, it is unable to focus on the mean value of time series and incapable to predict the intense. The researchers began to concentrate on machine learning parametric models due to the limitations in parametric model [9]. The nonparametric model depends on the training data to establish parameter numbers and structural model[10]. The machine learning approach called K-Nearest Neighbor (KNN) predicts the traffic stream by searching the closer neighbor matching to the present data from the traditional database. The Support Vector Machine (SVM) approach minimizes the structural risk and has advantages in high dimensional data and minute samples [11]. The Bayesian Network (BN) approach

detects the issues in partial data, which is related to the message transferring mechanism [12]. Similarly, the deep learning approaches are also applied in these fields like time series production, recognition and translation, etc. because of the ability in data mining [13]. However, deep learning approach is difficult to use and preprocessing the output and input series in high dimensional data is very slow. The other deep learning-based approach is Convolution Neural Networks (CNN) that is used to extract the significant data from the raw data [14] [15][16]. The hybrid network attains promising results than simple networks due to its capability in mining the attributes from the input traffic data [17]. Therefore, a hybrid networks method shows better results in predicting the traffic congestion. The traffic congestion prediction with hybrid network contains very few works due to the shortage of high-quality traffic congestion data.

The major contribution of the suggested traffic congestion prediction is given here:

- To design a new model to predict the traffic congestion by employing the traffic dataset with new hybrid algorithm for feature selection and hybrid architecture for prediction with congestion index computation by fuzzy interference system.
- To select the weighted features with HJHHO algorithm by optimizing the weight and selecting the features from the pre-processed data through minimizing the variance of features.
- To evaluate a proposed IDCNR prediction with Deformable Convolution Network (DCN) and Recurrent Neural Network (RNN) architecture by optimizing hidden neuron count of DCN, learning rate of DCN, epoch count of DCN and further, hidden neuron counts of RNN using HJHHO algorithm. The objective function of IDCNR is to minimize the MAE and RMSE.
- To compute congestion index computation by considering the inputs as the speed reduction rate, very low speed rate and volume to capacity rate by the well-performing fuzzy interference system.
- To determine the performance of the suggested model with other existing approaches by evaluating different measures to examine the outcomes of the proposed model.

The other sections of this paper are illustrated here. Section II represents the related works. Section III denotes the architectural view of optimal traffic congestion prediction with new hybrid heuristic algorithm with fuzzy logic system. Section IV denotes the enhanced traffic congestion prediction with new optimized weighted feature selection. Section V depicts the optimal traffic congestion prediction by hybrid deep learning with traffic congestion index computation by fuzzy model. Section VI denotes the result and discussion. Section VII indicates the conclusion.

II. LITERATURE SURVEY

A. Related Work

In 2018, Tseng et al. [18] have applied Apache storm in real time traffic congestion prediction scheme by analyzing various data such as rainfall volume, road density and traffic incidents. The new SVM-based Real-Time Highway Traffic Congestion Prediction (SRHTCP) method has gathered the traffic incident data reported in roadways from the Taiwan police broadcasting service and the weather reports were collected from Taiwan Central Weather Bureau. Here, the fuzzy set theory was used to estimate the traffic congestion in real time with regarding rainfall volume, road density, and road traffic incidents. The road speed for next timeline was predicted with SRHTCP method by analyzing the weather data and traffic flow data. The suggested SRHTCP scheme has achieved better accuracy prediction than other prediction models based on weighted exponential moving average approach.

In 2020, Ranjan et al. [19] have proposed three techniques to predict the traffic congestion like, an effective and low-cost data attainment scheme by captivating a snap of traffic clogging from the online open source web service TOPIS and a hybrid deep learning approach to predict the traffic congestion level by extracting the temporal and spatial information. The relationship of temporal and spatial for predicting the traffic congestion level was analyzed efficiently and effectively by the proposed model. This suggested model effectively has outperformed other Deep Neural Network (DNN) models.

In 2020, Shin et al. [20] have proposed a deep learning approach to predict the traffic congestion and in addition to correct the missing spatial and temporal values. The proposed model was pre-processed with outlier removal method using the Median Absolute Deviation (MAD) of traffic data. Therefore, the spatial and temporal values were corrected using spatial and temporal trends and pattern data. The proposed prediction model was combined with LSTM for learning the data in time series. The suggested model was compared over other existing model and has achieved better prediction results.

In 2019, Zhao et al. [21] have proposed an optimized GRU architecture to predict the speed of trucks on urban roads under non-periodic congested situations. The driving data of struck in Beijing Road was gathered. To get rid of the unwanted data, the pre-processing and screening of data were approached, and then, sequence of traffic speed was extracted. The learning rate was not adjusted by the Suitable Development Goals (SGD) weight optimization algorithm. The weight was optimized by Adadelta, Rmsprop and Adam in this proposed GRU model. The accuracy of the suggested model was verified based on four scenarios, such as accident, workday, rainy and weekend.

In 2020, Zaki et al. [22] have suggested a new Hidden Markov Model to determine the traffic stages in two dimensional (2D) spaces during the high peak hours. The proposed model has captured variance in traffic pattern data using the contrast and mean speed. The proposed model has enhanced the prediction error than other neuro fuzzy and HMM approaches. In 2019, Wena et al. [23] have suggested a Hybrid Temporal Association Rules Mining mode to predict the traffic clogging. The traffic states were predicted using DBSCAN algorithm, which has suitable rules in analyzing the traffic congestion in road ways. The temporal associated rules in traffic states were extracted using genetic algorithm based temporal association rules mining algorithm. The classification process was used to predict the traffic congestion level. The prediction and the stimulation tests were studied in different sizes of road ways. The stimulation results have determined that the suggested model has predicted the traffic congestion with high precision.

In 2016, Li et al. [24] have proposed an adaptive real time prediction model. This scheme has comprised a traffic pattern recognition algorithm related to an adaptive threshold calibration method, the adaptive K-means clustering and a 2D speed prediction method. The patterns from traffic data were obtained by the adaptive K-means clustering. The adaptive threshold calibration method was applied to recognize the traffic congestion and prediction. The obtained result has shown that the adaptive K-means clustering has recognized the traffic pattern better than Gaussian Mixture method. The proposed model has performed well in real time application of traffic congestion prediction and gained better accuracy.

In 2019, Song et al. [25] have applied the k-means clustering algorithm to classify the spatio-temporel distribution of congested roadways. Then, the spatiotemporal features were mined to extract the potential factors using geo-detector. The congested patterns were selected from six inter- regional and intra-regional roads on weekdays. The public properties like tourist spots, hospitals and green spots were often congested in off peak and peak hours. The result has suggested that the roads built-in high-density areas could reduce the repeated trips in center of the city. The land utilizing plan has involved with a detailed design of the environment to enhance various travel approaches in order to reduce the traffic congestion and increase the effectiveness of traffic. More advanced techniques were applied in land use plan and traffic congestion based on multi source real time data.

B. Problem Statement

Multi source data are collected and evaluated to predict and prevent the traffic congestion and traffic stream in transportation system. The training of all this data is still challenging. Various studies were undergone to handle this type of huge road data, which prevented the real time traffic congestion by using different techniques. Table I shows the features and challenges of traditional congestion prediction on traffic flow methods. SVM [18] returns better prediction accuracy and the car speed of the proceeding time period is predicted by the SRHTCP model. But, used open datasets are not verified using t-test technique. CNN and LSTM [19] train the image data using a large resolution on a smaller resource and better performance is achieved with respect to the computing time. Still, it does not enhance the performance of the model by including external factors such as weather information for every road. LSTM [20] solves the long term dependency problem and time-series features associated with the traffic data are learnt. Yet, it does not enhance the accuracy in the urban areas and low-speed regions. HMM [21] enhances the recovery vehicle count in a specific stretch of road at particular times and also supports the enhancement strategies and the traffic management on the longer term like ramp metering. But, the technique of contrast is not verified on extra datasets. GRU [22] offers an efficient information service for truck drivers and simultaneously meets the efficiency and accuracy of matching. Still, the applicability is not considered for the real traffic systems. GA [23] minimizes the correlation in the environments and also predicts the traffic congestion with a low error. Yet, it does not automatically choose the best parameter values. Adaptive K-means cluster [24] is insensitive for the actual congestion probability and also enhances the prediction performance in the case of real time. But it does not consider the external influencing factors like mega-event and weather for enhancing the traffic mode recognition. Mining technique [25] improves the accessibility for distinct travel modes and also identifies the potential urban form factors and the traffic congestion hotspots. Still, it does not consider the useful knowledge for the improved decision making using novel approaches. Thus, it is necessary to introduce novel deep learning methods for predicting the congestion in the traffic flow management system in order to reduce the error and enhance the accuracy of the overall system.

TABLE I. FEATURES AND CHALLENGES OF TRAFFIC CONGESTION PREDICTION MODELS

Author [citation]	Methodology	Features	Challenges
Tseng et al. [18]	SVM	The car speed of the proceeding time period is predicted by the SRHTCP model. It returns better prediction accuracy.	The used open datasets are not verified using t-test technique.
Ranjan et al. [19]	CNN and LSTM	Better performance is achieved with respect to the computing time. The image data is trained using a large resolution on a smaller resource.	The performance of the model is not enhanced by including external factors such as weather information for every road.
Shin et al. [20]	LSTM	The time-series features associated with the traffic data are learnt. The long term dependency problem is solved.	The accuracy is not enhanced in the urban areas and low-speed regions.
Zhao et al. [21]	HMM	It supports the enhancement strategies and the traffic management on the longer term like ramp metering. The recovery vehicle count is enhanced in a specific stretch of road at particular times.	The technique of contrast is not verified on extra datasets.
Zaki et al. [22]	GRU	It simultaneously meets the efficiency and accuracy of matching. An efficient information service is offered for truck drivers.	The applicability is not considered for the real traffic systems.
Wena et al. [23]	GA	The traffic congestion is predicted with a low error. The correlation in the environments is minimized.	The best parameter values are not automatically chosen.
Li et al. [24]	Adaptive K-means cluster	The prediction performance in the case of real time is enhanced. It is insensitive for the actual congestion probability.	The external influencing factors like mega-event and weather are not considered for enhancing the traffic mode recognition.
Song et al. [25]	Mining technique	It identifies the potential urban form factors and the traffic congestion hotspots. The accessibility is improved for distinct travel modes.	The useful knowledge for the improved decision making is not considered using novel approaches.

III. ARCHITECTURAL VIEW OF OPTIMAL TRAFFIC CONGESTION PREDICTION WITH NEW HYBRID HEURISTIC ALGORITHM WITH FUZZY LOGIC SYSTEM

A. Proposed Model and Description

Traffic congestion is one of the most important issues in all over the world. The current infrastructure is unable to cope with new traffic applications. The small spaces and other construction activities influence the traffic congestion. Due to the cause of traffic blockage, the fuel costs and the travel time of employers and distributing workers is affected. Traffic congestion is defined as the transportation vehicles surpass the capacity of roadway in peak time. The congestion indicators are mostly used to evaluate the traffic congestions in the urban road routes. Millions of peoples are affected by traffic congestion. This also causes noise and air pollution in whole surroundings. The impact of traffic blocking can be associated to fuel price raise, environment related matters, and transits cost. Various studies and researchers were undergone to overcome the traffic congestion. The timely prediction of traffic congestion in real time can prevent the unnecessary blockage. Deep learning and machine learning approaches were implemented to predict the traffic congestion. Machine learning-based model is most popular than other non-parametric models. It analyses the traffic patterns with low restrictions and gives better prediction results. Deep learning approaches are discussed to predict the real time traffic congestion. The traffic data are huge data, which are difficult to train. In this case, various techniques were executed to train the

huge data volume and to enhance the prediction accuracy. The architectural diagram for the proposed traffic congestion prediction is given in Fig. 1.

The proposed traffic congestion prediction model covers five main phases (a) data collection (b) Pre-processing (c) Feature selection (d) Prediction and (e) Congestion index computation. The benchmark dataset is gathered from Radar traffic counts, which is publicly available. These datasets are pre-processed by three methods such as removal of bad data, organizing the raw data and filling the null values. In this pre-processing phase, the raw data are cleaned by eliminating the unwanted data and filling the missing data. The pre-processed data is inputted to the optimized weighted feature selection phase, where the weighted feature selection is enhanced by hybrid HJHHO algorithm by optimizing the weight and features. The selected features are given to the classification phase. The features are predicted with DCN and RNN architecture by optimizing the hyperparameters like hidden neuron count, learning rate and epoch count using HJHHO algorithm. As the prediction uses hybridization of two deep leading models with architecture optimization, the proposed model is termed as IDCRN. The main purpose of the proposed prediction is to minimize the error rate and maximize the prediction performance. The predicted parameters such as speed reduction rate, low speed rate, and volume to capacity rate are attained from IDCRN. Finally, the predicted parameters are inputted to fuzzy interference system for congestion index computation such as high, low, moderate and very high.

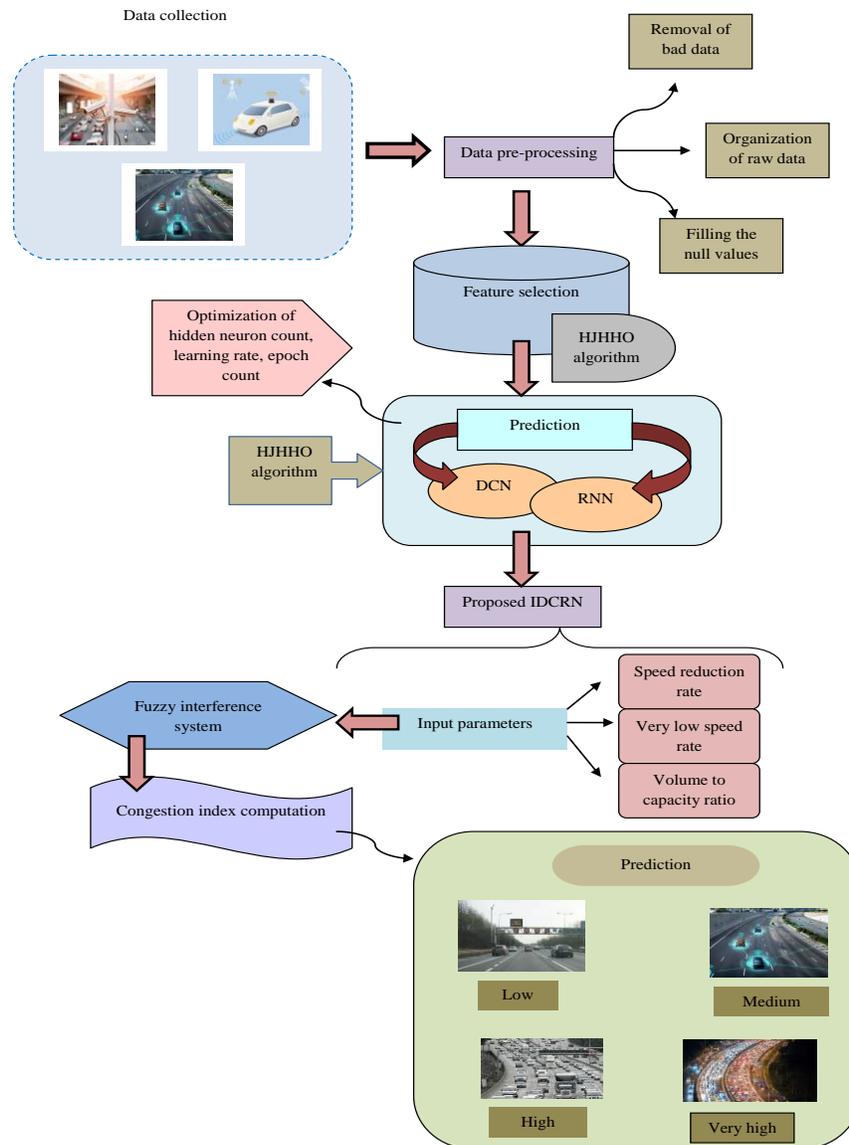


Fig. 1. Architectural Illustration of the Proposed Traffic Congestion Prediction Method with Congestion Index Computation.

B. Dataset Description

The benchmark dataset Radar Traffic counts are collected from “<https://data.austintexas.gov/Transportation-and-Mobility/Radar-Traffic-Counts/i626-g7ub>”- Access date: 24-12-2021. The traffic speed and count are gathered from various Wavetronix radar sensors, which is taken from the city of Austin. The dataset contains the hourly transportation of vehicles with 6.83 million of rows and 17 columns. The 70% of data is used for training and 30% of data is used for testing.

The collected input data is determined as T_f^{ip} , where $f = 1, 2, \dots, F$, here the term f is indicated as the total number of gathered dataset.

C. Data Pre-processing

The process of converting the collected raw data into suitable format is known as data pre-processing. The gathered

raw data are pre-processed to remove the unwanted data and to eliminate the noises before utilization. The pre-processing of raw data is essential for accurate precision analysis. In this proposed work, the pre-processing phase is performed with removal of bad data, organizing the raw data, and filling the null values.

Removing bad data: The input data T_f^{ip} is given to bad data removal process. Data cleaning or removal of bad data is a process to remove the unwanted data in the gathered dataset. The conversion of data from a structure or format to another is known as data transformation. It is essential to remove the bad data before utilization. In this process, the corrupted data, duplicate data, missing data are removed or fixed from the dataset. The presence of duplicate or unwanted data results in producing error in the prediction performance. The removal of bad data is denoted as T_f^{bad} .

Organizing the raw data: The data gathered are mostly unorganized or non-systematic, which are known as raw data. The deconstruction analyze method is used to manipulate or organize the data. The obtained raw data are in the form of recorded values and the systematic process of organizing them is referred as raw data organization. The organized data is denoted as T_f^{org} .

Filling the null values: The traffic data are usually affected in two different categories. First, the data are missed in certain time periods and locations. The entire data is necessary for the prediction and modelling of transportations. Second is the loss of statistical information. It causes violation of missing traffic patterns. The null values are filled with appropriate methods. When the missing values are to predict, the 0 or NA is used instead of the missing values. Filling the null values generates robust data models. Finally, the pre-processed data is denoted as T_f^{pre} and it is applied to feature selection phase.

IV. ENHANCED TRAFFIC CONGESTION PREDICTION WITH NEW OPTIMIZED WEIGHTED FEATURE SELECTION

A. Optimal Weighted Feature Selection

The pre-processed data T_f^{pre} are given to HJHHO algorithm for the feature selection. From the evaluated mining features, the optimized weighted features are obtained by multiplying with the optimized weight with the extracted features Fe_d^{sel} as given in Eq. (1).

$$Fe_{d^*}^{sel^*} = OP_d \times Fe_d^{sel} \quad (1)$$

In the above equation, the optimized weight is indicated as $OP_d, d = 1, 2, \dots, ND$, where ND indicates total number of features being considered and the term $Fe_{d^*}^{sel^*}$ is denoted as weighted optimal features. The weight is optimized with HJHHO algorithm in the range of $[0,1]$. The maximum iteration is fixed as 10. The main objective function of the weighted feature selection is to optimize the weight by minimizing the variance of the selected features. The number of optimal features selected from the extracted features is counted as 10.

$$objfn1 = \arg \max_{\{OP_d, Fe_d^{sel}\}} \left(\frac{1}{\text{var}} \right) \quad (2)$$

In Eq. (2), the term var refers to the variance, $objfn1$ is the objective function, the variance is defined as ‘‘a statistical measurement of the spread between numbers in a data set. More specifically, variance measures how far each number in the set is from the mean and thus from every other number in the set’’ as given in Eq. (3).

$$\text{Var} = \sigma^2 = \frac{\sum_{d=1}^r (y_d - \bar{y})^2}{r-1} \quad (3)$$

Here, the term y_d denotes the d^{th} data feature, \bar{y} represents the data features and r denotes the total number of data features. Consider, the optimal weighted features as $Fea_{d^*}^{sel^*} = Fe_1^{sel}, Fe_2^{sel}, Fe_3^{sel}, \dots, Fe_{ND^*}^{sel}$, where the total number of optimized weighted features are represented as ND^* .

The representation for optimal weighted feature selection is given in Fig. 2.

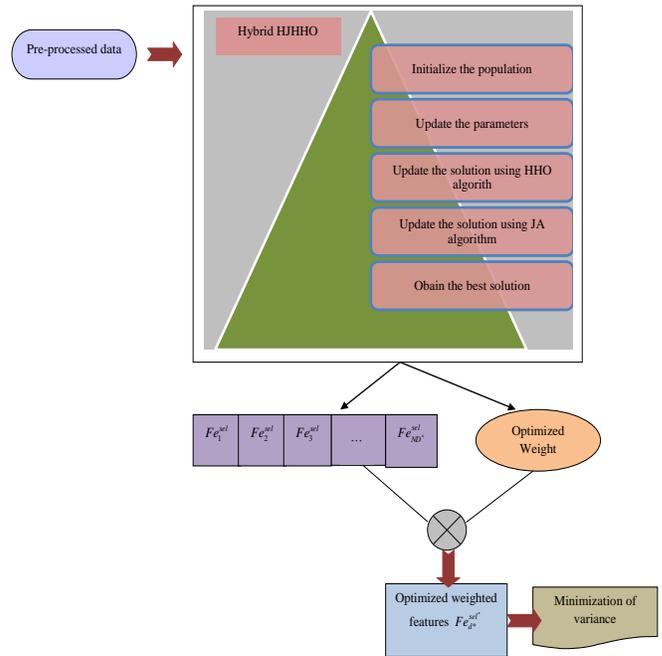


Fig. 2. Representation of Optimal Weighted Feature Selection.

B. Proposed HJHHO

The proposed HJHHO algorithm is used for selecting the optimized weighted feature selection from data by optimizing the selecting the features and weight and improving the hybrid based IDCRN prediction performance by optimizing the hidden neuron count of DCN, learning rate of DCN, epoch count of DCN and further, hidden neuron counts of RNN.

HHO [26] algorithm is flexible and used in various optimization problems to attain optimal solutions. It has high efficiency and predicted the failure probability using key factors. The exploration phase is maximized using compound agents. However, the HHO algorithm suffers from challenges like population diversity problem and local optima in high dimensional issues. To over these challenges in HHO, JA algorithm is used. JA [27] algorithm is easy and simple to implement and it contains few parameters in a single phase. It is applied to resolve various problems in optimization and finds optimal solutions within less computational time. The proposed hybrid HJHHO algorithm enhances the problems in optimization and more efficient than other optimization algorithms. If escaping energy $|B| \geq 1$, the position is updated using HHO through exploration phase, otherwise the position is updated using JA algorithm.

HHO algorithm is inspired by exploring and attacking behavior of Harris hawks. The HHO algorithm is gradient free population-based optimization technique, which can be proposed to all kinds of optimization problems. The exploration phase of HHO is utilized in HJHHO. Harris hawks can spot and track their prey with their powerful eyes. The prey is not visible often. The hawk waits in the desert spot for several hours to observe, monitor, and to detect the prey. The candidate solutions are denoted as Harris hawks and the best candidate solution is referred as the intentional prey. The hawk waits in a location for hours to detect the prey in terms of two strategies. The term r is denoted as the strategy of each perching, hawks perch depend on the position of their family members and the rabbit. If the condition $r < 0.5$, they perch on random trees is given in Eq. (4).

$$C(n+1) = \begin{cases} C_{rand}(n) - m_1 |C_{rand}(n) - 2m_2 C(n)| & r \geq 0.5 \\ \begin{cases} (C_{rabbit}(n) - C_k(n)) - \\ m_3(FS + m_4(WS - FS)) \end{cases} & r < 0.5 \end{cases} \quad (4)$$

In above equation, the term $C(n+1)$ is denoted as the position of hawks in next iteration n , $C_{rabbit}(n)$ denotes the position of rabbit, the term $C(n)$ is the current position of hawks, the random numbers are represented as m_1, m_2, m_3, m_4 , and r in $(0,1)$, the random numbers are upgraded in all iterations. The terms WS and FS are the lower and upper bounds of the variables, $C_{rand}(n)$ denotes the selected hawks from the present population, and C_k represents the average position of hawks from the present population.

At first, the solutions are generated depend up on its random location of hawks. Secondly, the location difference of the best candidate and the average position of the other hawks add a component depend on the ranges of variable. The scaling component m_3 increases the nature of rules once when m_4 get closer to the value 1 and similar distribution takes place. Based on this rule, a randomly scaled movement is added to FS . Then, a randomly scaled coefficient is considered to the component to offer more diversity trends and to explore various region of the attribute space. A simple rule is utilized to imitate the behavior of hawks. The average position of the hawks is represented in Eq. (5).

$$C_k(n) = \frac{1}{A} \sum_{o=1}^A C_o(n) \quad (5)$$

Here, the term $C_k(n)$ denotes the position of every hawk in iteration n and A refers to the total number of hawks. The average position is attained in various ways, but the simplest rule is utilized here. The HHO algorithm transfers the exploration phase to exploitation phase based on their escaping behavior. The escaping energy of the prey is denoted in Eq. (6).

$$B = 2B_h \left(1 - \frac{n}{N} \right) \quad (6)$$

Here, the term B denotes the escaping energy of the prey; the maximum iteration is represented as N and B_h is the initial stage of energy. The value B_h differs in interval $(-1, 1)$ at all iteration. If the B_h values decrease from 0 to -1 , that means the rabbit is failing, and when the B_h value maximizes from 0 to 1, that means the rabbit is strengthening. When the escaping energy of the rabbit is $|B| \geq 1$, the hawks search the rabbit in various locations, hence the exploration phase takes place and when $|B| < 1$ JA algorithm is used in exploitation phase.

JA algorithm [27] is very simple and used to resolve all optimization problems. The main purpose of JA algorithm is when a best solution is attained once for a certain problem, the results are obtained by eliminating the worst solutions immediately. JA algorithm obtains the optimal solution by ignoring the worst solution. It resolves the constrained and unconstrained optimization problems. Let $e(z)$ is denoted as the minimized major function. There are h number of intended variables ($w = 1, 2, \dots, h$) in all iteration n , o represents the candidate solution. The best candidate attains the best value of $e(z)$ ($e(z)_{best}$) in all candidate solution and further, the worst solution attains the worst value of $e(z)$ ($e(z)_{worst}$) in whole candidate solution. If $Z_{k,d,v}$ is the given value of k^{th} candidate in iteration n , the modified value is denoted in Eq. (7).

$$Z'_{k,d,v} = Z_{k,d,v} + m1_{k,v} \left(Z_{k,best,v} - |Z_{k,d,v}| \right) - m2_{k,v} \left(Z_{k,worst,v} - |Z_{k,d,v}| \right) \quad (7)$$

In above equation, the term $Z_{k,best,v}$ represents best candidate of variable k and the term $Z_{k,worst,v}$ is denotes the worst candidate of variable k . The term $Z'_{k,d,v}$ is the upgraded value of the terms $m1_{k,v}$, $Z_{k,d,v}$, and $Z_{k,d,v}$ are the two number of k^{th} candidate, in iteration at range $[0,1]$. The term $m1_{k,v} \left(Z_{k,best,v} - |Z_{k,d,v}| \right)$ indicates the inclination of the solutions to neglect the worst solution and the term $m2_{k,v} \left(Z_{k,worst,v} - |Z_{k,d,v}| \right)$ denotes the tendency to avoid the worst solution. At the end of iteration, the attained functional values are given to the next iteration. The pseudo code for designed HJHHO algorithm is given in algorithm 1.

Algorithm 1: Designed HJHHO

```

initialize of population
calculate the fitness of hawks
update the position of the hawks using Eq. (4)
update the position of rabbit using Eq. (4)
Determine the parameters  $B$  escaping energy
If  $|B| \geq 1$ 
    Update the solution using HHO
    update the position using exploration phase using Eq. (5)
else
    Update the solution using JA
    update the position using Eq. (7)
end for
    
```

The flowchart of designed HJHHO algorithm is given in Fig. 3.

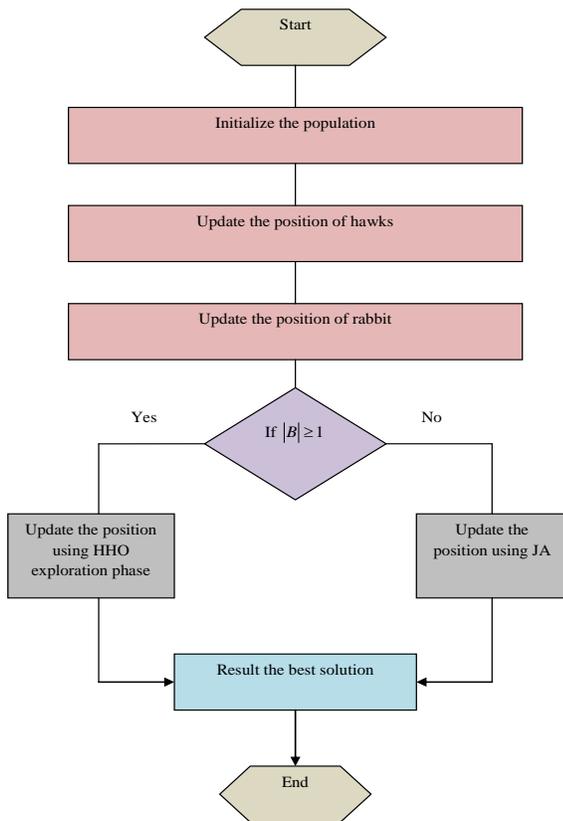


Fig. 3. The Flowchart for Designed HJHHO Algorithm.

V. OPTIMAL TRAFFIC CONGESTION PREDICTION BY HYBRID DEEP LEARNING WITH TRAFFIC CONGESTION INDEX COMPUTATION BY FUZZY MODEL

A. Proposed IDCRN

The proposed IDCRN is a hybridization of improved DCN and RNN, which is used to predict the traffic congestion parameters in the proposed mode. The main function of IDCRN is to predict the data by optimizing hidden neuron count of DCN, learning rate of DCN, epoch count of DCN and further, hidden neuron counts of RNN by HJHHO algorithm. The proposed IDCRN minimizes the MAE and RMSE measures, thus reducing the error rate.

DCN [28] architecture is used to extract the deep features and predict the data. The offset vector for each sampling is introduced with DCN. However, generalization ability of DCN can be reduced in some extent. In addition, the RNN are used in time series prediction as it retains information about previous input and it is capable to remember all information throughout time. However, the computation of RNN is slow and it is difficult to train. To overcome the drawbacks in CNN and RNN, the hybrid IDCRN model is introduced. The parameters related to the traffic congestion index are predicted using IDCRN.

DCN are used to extract the input feature maps, where the field of offset is calculated with convolution networks. The offset obtained from the additional convolution network is inputted to the original convolution network. The present location of the random sampling is realized by the kernels and the location is not inadequate for the standard grid. Each offset location is learned by the network rather than the convolutional kernels. The end-to-end spatial transformations are effectively and easily realized by DCN. To analyze the 2D convolutions, the regular grid N is used to model the input feature map y and the addition of sample values is weighed by a . The dilation and size of receptive fields are defined with the grid N . The convolution kernel 3×3 with dilation 1 and $N = \{(-1, -1), (-1, 0), \dots, (0, 1), (1, 1)\}$ is taken. In each location r_o , the output feature map x is given as represented in Eq. (8).

$$x(r_o) = \sum_{r_s} a r_s \cdot y(r_o + r_s) \quad (8)$$

Here, the term r_s computes the location in N . In DCN, the regular grid N is enhanced with the offsets $\{\Delta r_o | n1, \dots, N\}$, where $S = |N|$. The modified equation is given in Eq. (9).

$$x(r_o) = \sum_{r_s} a r_s \cdot y(r_o + r_s + \Delta r_s) \quad (9)$$

Here, the term $r_s + \Delta r_s$ represents the irregular location of offset, in which, Δr_s is naturally fractional. The bilinear interpolation is given in Eq. (10).

$$y(r) = \sum_t K(t, r) \cdot y(t) \quad (10)$$

Here, the term $r = r_o + r_s + \Delta r_s$ indicates the fractional location, t denotes all spatial location present in feature map y , the term $K(\cdot, \cdot)$ denotes the bilinear interpolation of kernel. The bilinear interpolation defines the linear interpolation in Y direction and X direction. The term K is divided into 2D dimensional kernel is referred in Eq. (11).

$$K(t, r) = k(t_y, r_y) \cdot k(t_x, r_x) \quad (11)$$

Hence, the weighted feature data are predicted using DCN architecture.

RNN [is another type of Artificial Neural Networks (ANN), which is applied in prediction of sequential traffic data by using previous traffic data. RNN generates the sequential information and each sequence of the element executes the similar task. The information is recorded in the memory of RNN. The input sequence of RNN is represented as C , which analyses the state t_{xx} at time xx . The different function v is applied with the previous state t_{xx-1} . The term t_{xx-1} does not contain the previous information. The RNN weight parameters are denoted as B , D and Y respectively. The RNN formula is given in Eq. (12) and Eq. (13).

$$t_{xx-1} = t_{xx-1} * Y + c_{xx} * B \quad (12)$$

$$o_{xx} = t_{xx} * D \quad (13)$$

The RNN model reduces the loss function and the errors in the predictions are compared with the actual values. The loss function is denoted as w , with regards the output state at given time xx . The correlation of gradient parameters is given here.

$$\frac{\partial w}{\partial_{xx-1}} = \frac{\partial w}{\partial_{xx}} Y \quad (14)$$

$$\frac{\partial w}{\partial B} = \sum_{xx=0}^E \frac{\partial w}{\partial_{xx}} c_{xx} \quad (15)$$

$$\frac{\partial w}{\partial Y} = \sum_{xx=0}^e \frac{\partial w}{\partial_{xx}} t_{xx-1} \quad (16)$$

Finally, the prediction data outcomes are attained from RNN. The optimized weighted features are inputted in DCN and RNN and the parameters such as speed reduction rate, volume to capacity rate and very low speed rate are predicted. The average from both the prediction models is utilized to perform the traffic congestion prediction.

B. Objective Function for IDCRN-based Parameter Prediction

The proposed IDCRN-based parameter prediction approach is used to predict the traffic congestion. The data is predicted using DCN and RNN. The parameters such as volume, time and speed of vehicles are inputted to IDCRN. The obtained output parameters such as volume to capacity rate, speed reduction rate and very low speed rate from IDCRN are subjected to fuzzy interference system to predict the congestion index computation in traffic flow. The main aim of the proposed prediction model is to optimize the hidden neuron count of DCN, learning rate of DCN, epoch count of DCN and further, hidden neuron count of RNN by minimizing the MAE and RMSE is given in Eq. (17).

$$objfn2 = \arg \max_{\{HN, LR, EN, HRN\}} (MAE + RMSE) \quad (17)$$

In above equation, the term HN denote the hidden neuron count of DCN that is ranging from $[5, 255]$, LR denotes the learning rate of DCN lies in the range of $[0.01, 0.99]$, EN represents the epoch count of DCN with the bounding limit of $[2, 20]$ and further, the hidden neuron count of RNN HRN ranging from $[5, 255]$. The term MAE indicates Mean Absolute Error and $RMSE$ denotes Root Mean Square Error. MAE is known as the “measure of errors between paired observations expressing the same phenomenon” as shown in Eq. (18).

$$MAE = \frac{\sum_{o=1}^r |xx_o - yy_o|}{r} \quad (18)$$

RMSE “is one of the most commonly used measures for evaluating the quality of predictions. It shows how far predictions fall from measured true values using Euclidean distance” as shown in Eq. (19).

$$RMSE = \frac{\sqrt{\sum_{o=1}^r \|yy_o - xx_o\|^2}}{r} \quad (19)$$

In above equations, term xx is represented as actual value and yy is denoted as forecasted value. The congestion measures like speed reduction rate, volume to capacity rate and very low speed rate are inputted to fuzzy interference for congestion index computation, each of these parameters indicates the traffic flow rates accurately. The fuzzy interference gives less error rate. The architectural diagram for IDCRN-based prediction parameters is given in Fig. 4.

C. Congestion Index Computation by Fuzzy Interference System

The training of different congestion measures has individual advantages and drawbacks. The congestion is an incident which is caused by various factors and the efforts are incorporated in different measures. The traffic congestion measurement is significant to detect the passenger perception. The boundary between each passenger differs due to the travel situations. These limitations are considered and fuzzy interference system is incorporated to detect the indistinct boundary in a set, and to find the solution uncertainty problems. The values of input parameters are calculated, categorized into various groups, determined different stages of traffic congestion and then finally established the congestion index computation.

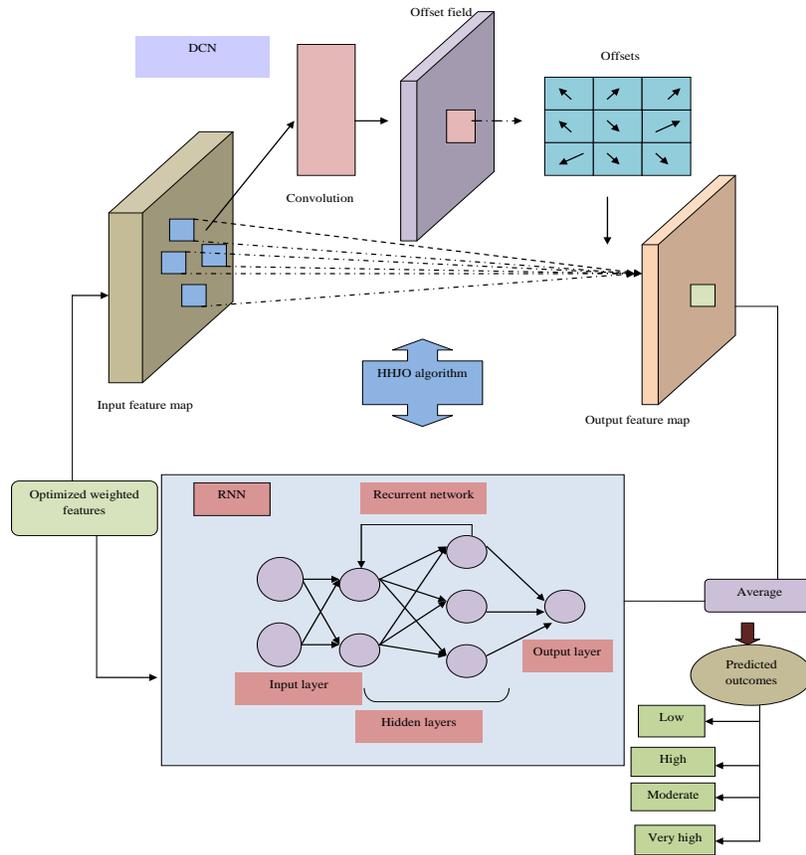


Fig. 4. Architectural diagram for IDCARN-based Parameter Prediction

Input parameters: The input parameters are observed from proposed IDCARN and then combined to form a single fuzzy measure. The three parameters such as speed reduction rate, volume to capacity ratio and very low speed rate are compared with traffic volume and traffic travel time to roadway capacity rate. These three parameters are calculated separately based on the gathered data and combined according to fuzzy interference system rule. The traffic condition is represented with these three input parameters, by varying the volume to capacity, average speed and the variation in speed.

Speed reduction rate: Speed reduction rate is used to minimize the speed of vehicles that causes traffic congestion. The congestion in various routes during peak and non-peak situation are compared with speed reduction rate as given in Eq. (20).

$$\text{Speed reduction rate} = \frac{(NPS - PS)}{NPS} \quad (20)$$

In above Eq. (20), the term NPS denotes the non-peak flow speed and PS denotes the peak flow speed. The speed obtained value ranges from [0, 1], where 1 is considered as the worst state when the peak flow speed is nearer to 0. The 0 is denoted as best state when the peak range is equal to or larger than non-peak flow speed.

Very low speed rate: It is defined as the amount of travelling time at very low speed when compared over the total travel period as given in Eq. (21).

$$\text{Very low speed rate} = \frac{TSD}{TT} \quad (21)$$

In Eq. (21), the term TSD is represented as times spend in delay and the term TT denotes total travel time. The obtained value ranges from [0, 1], where 0 denotes the best state with no delay and 1 denotes the worst state with delayed travel time. The delay is referred as the time travel speed, which must be lesser than 5km/hr.

Volume to capacity ratio: It calculates all traffic block during peak hour situations. The volume is computed by taking the vehicles unit per hour, when the capacity of the roadway is increased as given in Eq. (22).

$$\text{Volume to capacity ratio} = \frac{VP}{CR} \quad (22)$$

In Eq. (22), the term VP denotes the volume of vehicle in peak hour and CR represents the capacity of roadway. The obtained value ranges from [0, 3], where the value closer to 0 is the best condition, when the capacity to ratio is minimized and value 3 is the worst condition when the huge volume of

vehicles moving towards the roadway by comparing to its capacity.

Output parameters: The output process is termed as congestion index, which is grouped into four phases low, high, very high and moderate. The function condition is given in a scale range from 0 to 1, where 0 is denoted as good and 1 as bad. The congestion condition of the four phases is determined in terms of this scale value.

Rules: The task of compiling three inputs and evaluating the traffic congestion measures are executed with a fuzzy inference system. When the speed reduction rate is considered as S, the very low speed rate is denoted as R, the volume to capacity ratio is indicated as V, and the congestion is referred as the term C. Here, S, R, V and C denote the degree of congestion. In the following rule, IF part is known as antecedent and THEN part is known as consequent. The exemplary representation of the generated fuzzy rules is given below.

- IF the speed reduction rate is low, AND the volume to capacity ratio is low, AND the very low speed rate is low, THEN, the congestion is low.
- IF the speed reduction rate is high, the volume to capacity ratio is high, AND the very low speed rate is high, THEN, the congestion is very high.

In the same way, totally, 54 numbers of rules are considered, among them three intense conditions are appropriate for using the combination of three categories such as volume to capacity ratio, speed reduction rate and very low speed. Among these rules, 38 numbers of suitable rules are

evaluated in this process to get required output. In this way, the traffic congestion index is computed. The architectural illustration for congestion index computation with fuzzy inference system is given in Fig. 5.

D. Output of Congestion Index by Fuzzy Inference System

The output of congestion index by fuzzy inference system in terms of membership function is given in Fig. 6.

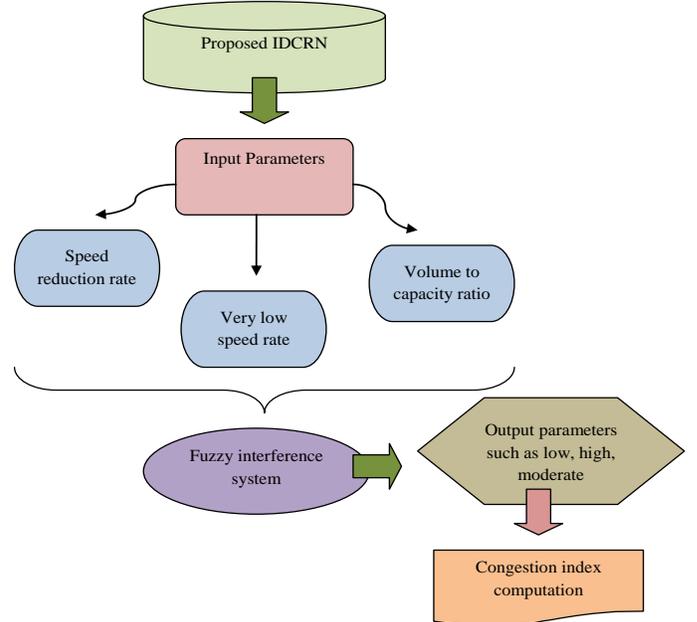


Fig. 5. Architectural Illustration of Congestion Index Computation with Fuzzy Inference System.

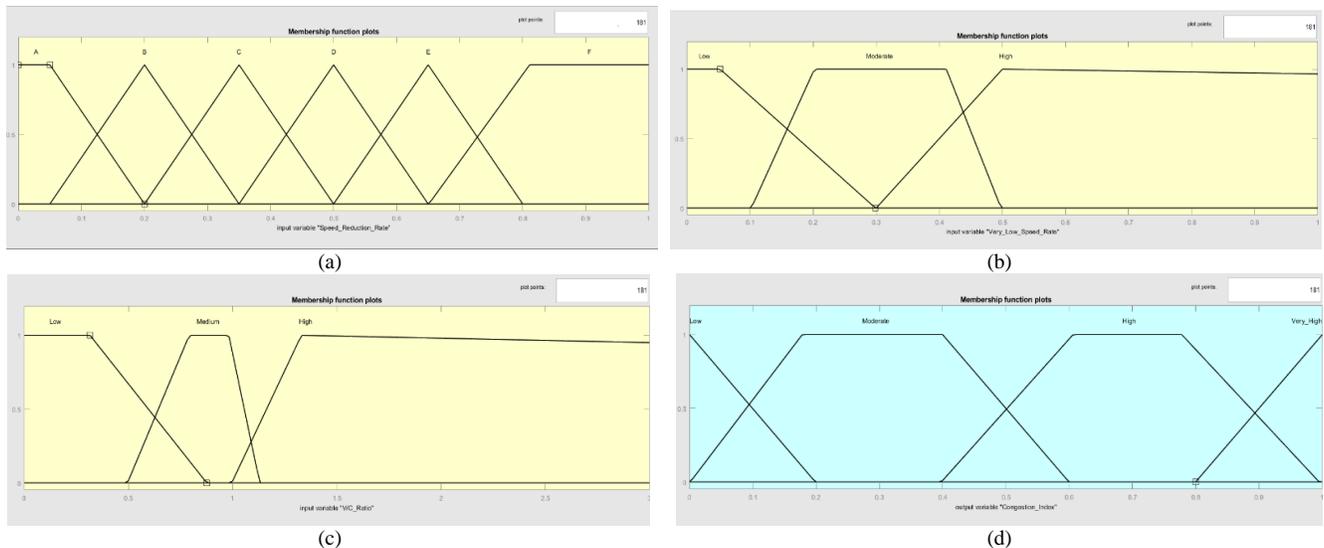


Fig. 6. Input and Output of Fuzzy Inference System for “(a) Speed Reduction Rate, (b) Very Low Speed Rate, (c) Volume to Capacity Ratio and (d) Final Congestion Index.

VI. RESULT AND DISCUSSION

A. Experimental Setup

The proposed model traffic congestion prediction was implemented in python and the experimental results were evaluated. The performance of the suggested model was compared over various existing models in terms of different measures. The suggested model was compared with different heuristic algorithms and prediction approaches. The maximum iteration was 10 and the number of populations was 10. The proposed is compared over various heuristic algorithms like GWO-IDCRN [29], WOA-IDCRN [30], HHO-IDCRN [26], JA-IDCRN [27] and prediction models like LSTM [31], CNN [32], RNN [33], and CNN-RNN [34].

B. Performance Metrics

The performance measures used for traffic congestion prediction are given here.

1) *L1-Norm* “is the sum of the magnitudes of the vectors in a space. It is the most natural way of measure distance between vectors that is the sum of absolute difference of the components of the vectors” as shown in Eq. (23).

$$L1 - Norm = \|yy\|_1 = \sum_{o=1}^{rs} |yy_o| \quad (23)$$

2) *L2-Norm* “is also known as the Euclidean norm. It is the shortest distance to go from one point to another” as shown in Eq. (24).

$$L2 - Norm = \|yy\|_2 = \sqrt{\sum_{o=1}^{rs} yy_o^2} \quad (24)$$

3) *L-infinity Norm* “is the vector space of essentially bounded measurable functions with the essential supreme norm and only the largest element has any effect” as shown in Eq. (25).

$$L - infinity\ norm = \|yy\|_\infty = \max_{1 \leq o \leq rs} |yy_o| \quad (25)$$

4) *MAE*: it is given in Eq. (18).

5) *Mean Absolute Scaled Error (MASE)* “is a measure of the accuracy of forecasts. It is the mean absolute error of the forecast values, divided by the mean absolute error of the in-sample one-step naive forecast” as shown in Eq. (26).

$$MASE = \frac{1}{RS} \sum_{o=1}^{rs} u_o \quad (26)$$

6) *Mean Effective Pressure (MEP)* “is a theoretical parameter used to measure the performance of an internal combustion engine” as shown in Eq. (27).

$$MEP = \frac{100\%}{rs} \sum_{o=1}^{rs} \frac{xx_o - yy_o}{xx_o} \quad (27)$$

7) *RMSE*: It is given in Eq. (19).

8) *Symmetric Mean Absolute Percentage Error (SMAPE)* “is used to measure the predictive accuracy of models” as shown in Eq. (28).

$$SMAPE = \frac{1}{rs} \sum_{o=1}^{rs} \frac{|xx_o - yy_o|}{(|xx_o| + |yy_o|) / 2} \quad (28)$$

C. Performance Analysis on Heuristic Algorithms

The performance analysis of the suggested model HJHHO-IDCRN algorithm is compared with other meta-heuristic algorithms by varying the learning percentage is given in Fig. 7. The MAE of proposed HJHHO-IDCRN algorithm is 1% higher than GWO-IDCRN, 3% higher than WOA, 1% higher than HHO-IDCRN and 4% higher than JA-IDCRN at learning percentage 60. The MEP measure had attained higher results over other heuristic algorithms, thus the proposed HJHHO-IDCRN algorithm is 3% superior to GWO-IDCRN, 2% superior to WOA-IDCRN, 4% superior to HHO-IDCRN and 2% superior to JA-IDCRN at learning percentage 40. The performance of the suggested algorithm has reduced the error rate and obtained high prediction accuracy than other existing algorithms.

D. Performance Measures on Prediction Models

The performance analysis of the proposed HJHHO-IDCRN algorithm is compared with other prediction models by varying learning percentage is given in Fig. 8. At learning percentage 55, the L-infinity form measures of proposed HJHHO-IDCRN algorithm is 12% higher than LSTM, 16% higher than CNN, 11% higher than RNN and 16% higher than CNN-RNN. The MAE measures of HJHHO-IDCRN algorithm are 2% is improved than LSTM, 1% improved than CNN, 3% improved than RNN and 5% improved than CNN-RNN at learning percentage 75. The MEP measures had obtained higher results over other heuristic algorithms, thus the proposed HJHHO-IDCRN algorithm is 8% superior to LSTM, 4% superior to CNN, 5% superior to RNN and 6% superior to CNN-RNN at learning percentage 85. The proposed algorithm has gained high prediction accuracy than other prediction models and reduced the error rate.

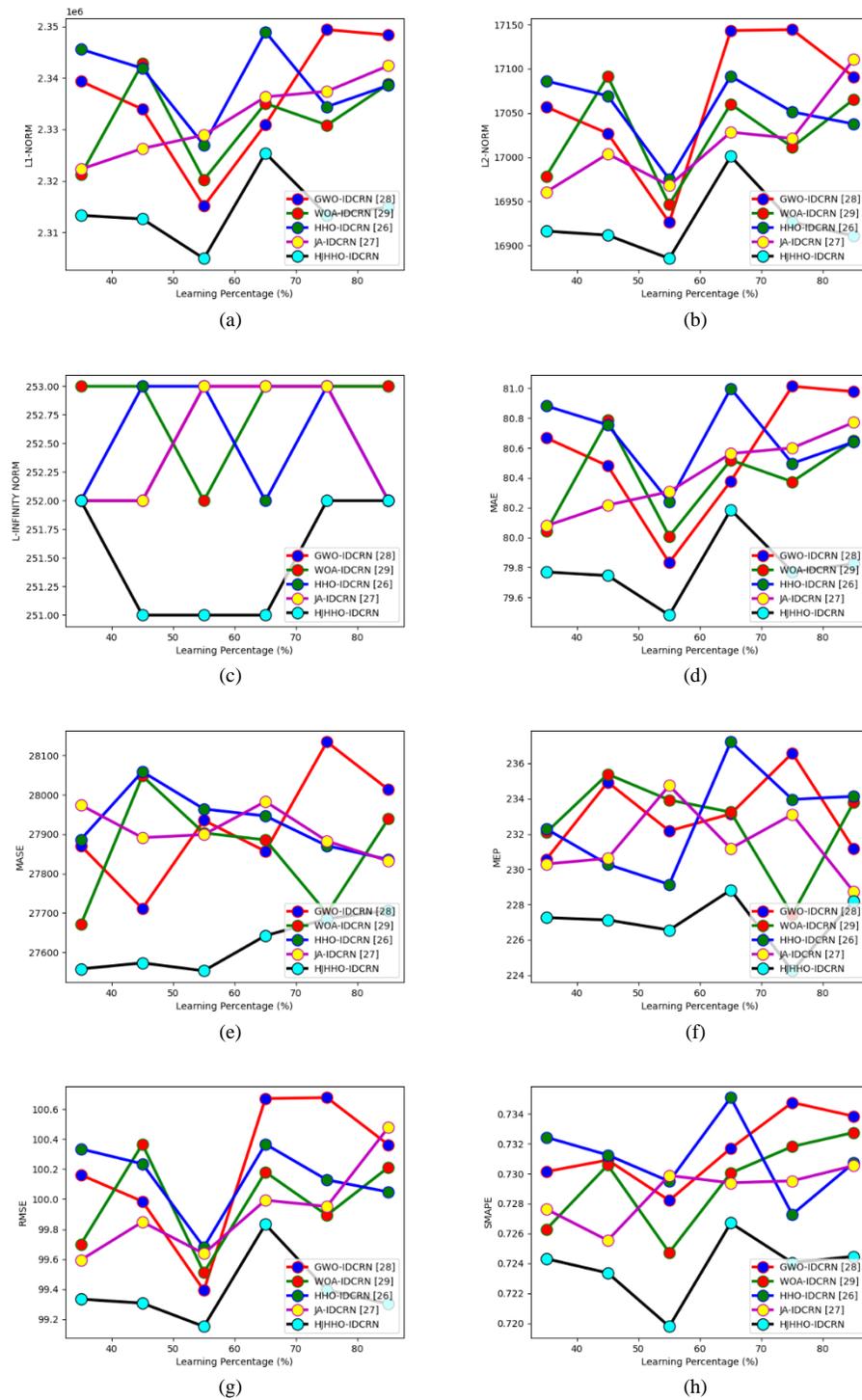


Fig. 7. Analysis of Traffic Congestion Prediction over other Heuristic Algorithms by Varying Learning Percentage in Terms of (a) L1Norm (b) L2-Norm (c) L-Infinity Norm (d) MAE (e) MASE (f) MEP (g) RMSE (h) SMAPE.

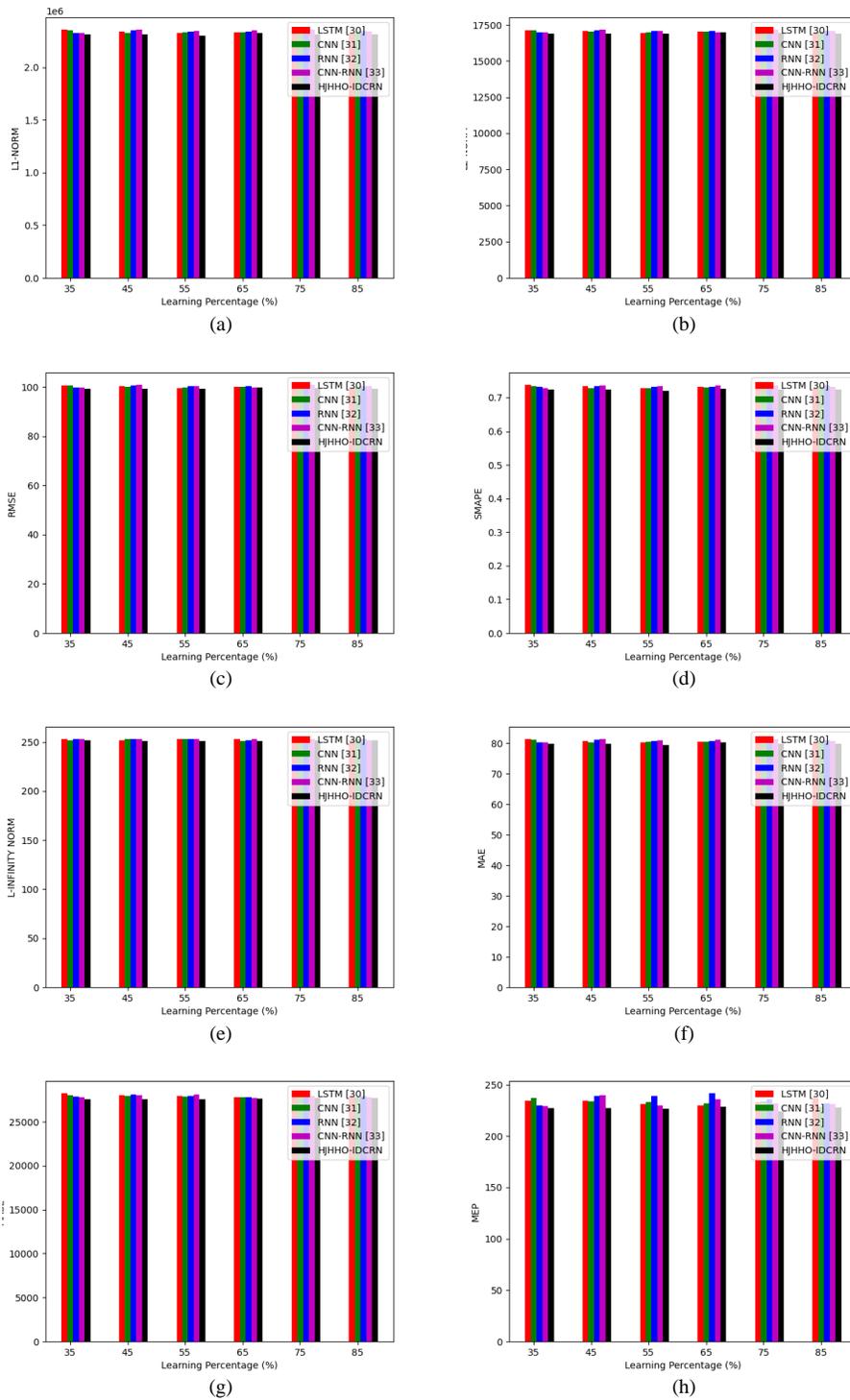


Fig. 8. Analysis of Traffic Congestion Prediction over other Prediction Models by Varying Learning Percentage in Terms of (a) L1Norm (b) L2-Norm (c) L-Infinity Norm (d) MAE (e) MASE (f) MEP (g) RMSE (h) SMAPE.

E. Overall Performance Analysis on Algorithms

The overall performance analysis of traffic congestion is evaluated with HJHHO-IDCRN is illustrated in Table II. The SMAPE measures of the proposed HJHHO-IDCRN algorithm is 5% higher than GWO-IDCRN, 3% higher than WOA-IDCRN, 1% higher than HHO-IDCRN and 6% higher than JA-IDCRN. The RMSE measures of proposed HJHHO-IDCRN algorithm is 15% superior to GWO-IDCRN, 11% superior to WOA-IDCRN, 14% superior to HHO-IDCRN and 9% superior to JA-IDCRN. The MAE measures of proposed HJHHO-IDCRN algorithm is 3% higher than GWO-IDCRN, 1% higher than WOA-IDCRN, 5% higher than HHO-IDCRN and 4% higher than JA-IDCRN. Therefore, the outcomes of the proposed HJHHO-IDCRN algorithm obtain better results than other heuristic algorithms while evaluating with all the measures.

F. Overall Performance Analysis on Prediction Models

The overall performance measures of the proposed HJHHO-IDCRN algorithm are compared with various prediction models and the prediction measures are given in Table III. The MASE measures of proposed HJHHO-IDCRN

algorithm is 2% better than LSTM, 4% better than CNN, 3% better than RNN and 6% better than CNN-RNN. The MEP measures of proposed HJHHO-IDCRN algorithm is 11% better than LSTM, 22% better than CNN, 16% better than RNN and 15% better than CNN-RNN. The MEP measures had gained higher results over other prediction models, thus the proposed HJHHO-IDCRN algorithm is 15% superior to LSTM, 22% superior to CNN, 15% superior to RNN and 16% superior to CNN-RNN. The suggested HJHHO-IDCRN algorithms give better prediction performance and reduce the error rate.

G. Results on Traffic Congestion Index

The obtained congestion parameter results are illustrated in Table IV. The three parameters capture the real condition of traffic congestion. The results have shown when a particular input parameter is high, then the other two parameters are on lower state. This state defines the moderate condition. When all the three parameters are on higher state, it results in high congestion result. In other hand, when all three parameters are in lower state, the congestion index results in very low values. The proposed model gives accurate results in traffic congestion prediction while comparing to other existing methods.

TABLE II. OVERALL PERFORMANCE OF TRAFFIC CONGESTION PREDICTION OVER OTHER META-HEURISTIC ALGORITHMS

Algorithms	GWO-IDCRN [29]	WOA-IDCRN[30]	HHO-IDCRN[26]	JA-IDCRN[27]	HJHHO-IDCRN
MEP	236.5818	227.4271	233.9635	233.1066	224.2438
SMAPE	0.734767	0.731819	0.727276	0.729521	0.724051
MASE	28134.79	27699.09	27870.95	27882.89	27685.07
MAE	81.01386	80.37438	80.49721	80.60052	79.77148
RMSE	100.6757	99.89368	100.1283	99.95179	99.39565
L1-NORM	2349402	2330857	2334419	2337415	2313373
L2-NORM	17144.45	17011.28	17051.24	17021.18	16926.47
L-INFINITY NORM	253	253	253	253	252

TABLE III. OVERALL PERFORMANCE OF TRAFFIC CONGESTION PREDICTION OVER OTHER PREDICTION MODELS

Algorithms	LSTM[31]	CNN[32]	RNN[33]	CNN-RNN[34]	HJHHO-IDCRN
MEP	233.3224	233.7626	235.8281	232.0925	224.2438
SMAPE	0.728538	0.732675	0.732912	0.735904	0.724051
MASE	27856.46	28007	28206.31	27931.84	27685.07
MAE	80.18579	80.69983	81.0989	81.31279	79.77148
RMSE	99.61117	100.3257	100.5825	100.767	99.39565
L1-NORM	2325388	2340295	2351868	2358071	2313373
L2-NORM	16963.17	17084.86	17128.58	17159.99	16926.47
L-INFINITY NORM	252	253	252	253	252

TABLE IV. RESULTS ON TRAFFIC CONGESTION INDEX

Speed reduction rate	Volume to capacity ratio	Very low speed rate	Congestion index
0.669068	0.302595	0.282547	0.467693
0.669252	0.304988	0.294872	0.412533
0.68714	0.448816	0.294838	0.531245
0.671981	0.319807	0.239569	0.765443
0.662298	0.740872	0.361626	0.678416
0.66062	0.416873	0.537614	0.504879
0.658038	0.174094	0.277427	0.43729
0.691069	0.394865	0.398412	0.268478
0.667211	0.122501	0.398846	0.309538
0.671858	0.359507	0.247754	0.60257
0.666126	0.246816	0.274632	0.610337
0.683519	0.244399	0.281251	0.586398
1	0.546707	1	0.380909
0.654448	0.38242	0.952603	0.450354
0.704439	0.276988	0.366578	0.343328
0.668823	0.300416	0.247549	0.348988
0.680675	0.393189	0.398136	0.440765
0.657003	0.98637	0.511805	0.254022

VII. CONCLUSION

In this suggested work, a hybrid-based deep learning approach with congestion index computation by fuzzy model was implemented to predict the traffic congestion. The traffic data was gathered from a publicly available database. The dataset was pre-processed with three phases to remove the unwanted data and to fill the missing data by removing bad data, organizing the raw data and filling the null values techniques. The weighted feature selection of data was performed with proposed HJHHO algorithm by optimizing the weight and minimizing the variance. The prediction phase was done with proposed IDCRN. The main purpose of prediction phase was to minimize the MAE and EMSE measures by optimizing the hidden neuron count of DCN, learning rate of DCN, epoch count of DCN and further, hidden neuron counts of RNN. The congestion parameters such as speed reduction, volume to capacity ratio and very low speed rate obtained from IDCRN were subjected to fuzzy interference system for congestion index computation. Finally, the proposed HJHHO-IDCRN algorithm has improved the performance of prediction. The performance analysis has shown the MAE measures of HJHHO-IDCRN algorithm was 5% higher than GWO-IDCRN, 3% higher than WOA-IDCRN, 1% higher than HHO-IDCRN and 6% higher than JA-IDCRN. Thus, the suggested HJHHO-IDCRN algorithm has attained better prediction in traffic congestion. In future works, the proposed model will be run over different datasets of different road networks and furthermore it will be deployed in route planning solutions.

REFERENCES

[1] C. Onyeneke, "Modeling the Effects of Traffic Congestion on Economic Activities - Accidents, Fatalities and Casualties," *Biomed. Stat. Informatics*, vol. 3, no. 2, p. 7, 2018, doi: 10.11648/j.bsi.20180302.11.

[2] Z. Huang, J. Xia, F. Li, Z. Li, and Q. Li, "A Peak Traffic Congestion Prediction Method Based on Bus Driving Time," *entropy*, vol. 21, no. 709, pp. 1–18, 2019.

[3] M. Chen, X. Yu, and Y. Liu, "PCNN: Deep Convolutional Networks for Short-Term Traffic Congestion Prediction," *IEEE Trans. Intell. Transp. Syst.*, vol. 19, no. 11, pp. 3550–3559, 2018, doi: 10.1109/TITS.2018.2835523.

[4] A. Elfar, A. Talebpour, and H. S. Mahmassani, "Machine Learning Approach to Short-Term Traffic Congestion Prediction in a Connected Environment," *Transp. Res. Rec.*, vol. 2672, no. 45, pp. 185–195, 2018, doi: 10.1177/0361198118795010.

[5] X. Ma, H. Yu, Y. Wang, and Y. Wang, "Large-scale transportation network congestion evolution prediction using deep learning theory," *PLoS One*, vol. 10, no. 3, pp. 1–17, 2015, doi: 10.1371/journal.pone.0119044.

[6] S. Zhang, Y. Yao, J. Hu, Y. Zhao, S. Li, and J. Hu, "Deep autoencoder neural networks for short-term traffic congestion prediction of transportation networks," *Sensors (Switzerland)*, vol. 19, no. 10, pp. 1–19, 2019, doi: 10.3390/s19102229.

[7] S. Ye, "Research on Urban Road Traffic Congestion Charging Based on Sustainable Development," *Phys. Procedia*, vol. 24, pp. 1567–1572, 2012, doi: 10.1016/j.phpro.2012.02.231.

[8] C. Wang, M. A. Quddus, and S. G. Ison, "Impact of traffic congestion on road accidents: A spatial analysis of the M25 motorway in England," *Accid. Anal. Prev.*, vol. 41, no. 4, pp. 798–808, 2009, doi: 10.1016/j.aap.2009.04.002.

[9] J. Zhao, Y. Gao, Z. Bai, H. Wang, and S. Lu, "Traffic speed prediction under non-recurrent congestion: based on lstm method and beidou navigation satellite system data," *IEEE Intell. Transp. Syst. Mag.*, vol. 11, no. 2, pp. 70–81, 2019, doi: 10.1109/MITS.2019.2903431.

[10] S. Sun, J. Chen, and J. Sun, "Traffic congestion prediction based on GPS trajectory data," *Int. J. Distrib. Sens. Networks*, vol. 15, no. 5, 2019, doi: 10.1177/1550147719847440.

[11] M. Aftabuzzaman, G. Currie, and M. Sarvi, "Evaluating the Congestion Relief Impacts of Public Transport in Monetary Terms," *J. Public Transp.*, vol. 13, no. 1, pp. 1–24, 2010, doi: 10.5038/2375-0901.13.1.1.

[12] M. Bani Younes and A. Boukerche, "A performance evaluation of an efficient traffic congestion detection protocol (ECODE) for intelligent

- transportation systems,” *Ad Hoc Networks*, vol. 24, no. PA, pp. 317–336, 2015, doi: 10.1016/j.adhoc.2014.09.005.
- [13] X. Yang, Yuanfeng and Cui, Z. and Wu, J. and Zhang, G. and Xian, “Fuzzy c-means clustering and opposition-based reinforcement learning for traffic congestion identification,” *J. Inf. Comput. Sci.*, vol. 9, pp. 2441–2450, 2012.
- [14] Y. Gu and L. Xie, “Fast sensitivity analysis approach to assessing congestion induced wind curtailment,” *IEEE Trans. Power Syst.*, vol. 29, no. 1, pp. 101–110, 2014, doi: 10.1109/TPWRS.2013.2282286.
- [15] H. Shankar, P. L. N. Raju, and K. R. M. Rao, “Multi Model Criteria for the Estimation of Road Traffic Congestion from Traffic Flow Information Based on Fuzzy Logic,” *J. Transp. Technol.*, vol. 02, no. 01, pp. 50–62, 2012, doi: 10.4236/jts.2012.21006.
- [16] Y. Zhang, N. Ye, R. Wang, and R. Malekian, “A method for traffic congestion clustering judgment based on grey relational analysis,” *ISPRS Int. J. Geo-Information*, vol. 5, no. 5, pp. 1–15, 2016, doi: 10.3390/ijgi5050071.
- [17] J. F. W. Zaki, A. M. T. Ali-Eldin, S. E. Hussein, S. F. Saraya, and F. F. Areed, “Time aware hybrid hidden markov models for traffic congestion prediction,” *Int. J. Electr. Eng. Informatics*, vol. 11, no. 1, pp. 1–17, 2019, doi: 10.15676/ijeei.2019.11.1.1.
- [18] F. H. Tseng, J. H. Hsueh, C. W. Tseng, Y. T. Yang, H. C. Chao, and L. Der Chou, “Congestion prediction with big data for real-time highway traffic,” *IEEE Access*, vol. 6, pp. 57311–57323, 2018, doi: 10.1109/ACCESS.2018.2873569.
- [19] N. Ranjan, S. Bhandari, H. P. Zhao, H. Kim, and P. Khan, “City-wide traffic congestion prediction based on CNN, LSTM and transpose CNN,” *IEEE Access*, vol. 8, pp. 81606–81620, 2020, doi: 10.1109/ACCESS.2020.2991462.
- [20] D. H. Shin, K. Chung, and R. C. Park, “Prediction of Traffic Congestion Based on LSTM through Correction of Missing Temporal and Spatial Data,” *IEEE Access*, vol. 8, pp. 150784–150796, 2020, doi: 10.1109/ACCESS.2020.3016469.
- [21] J. Zhao et al., “Truck traffic speed prediction under non-recurrent congestion: Based on optimized deep learning algorithms and GPS Data,” *IEEE Access*, vol. 7, pp. 9116–9127, 2019, doi: 10.1109/ACCESS.2018.2890414.
- [22] J. F. Zaki, A. Ali-Eldin, S. E. Hussein, S. F. Saraya, and F. F. Areed, “Traffic congestion prediction based on Hidden Markov Models and contrast measure,” *Ain Shams Eng. J.*, vol. 11, no. 3, pp. 535–551, 2019, doi: 10.1016/j.asej.2019.10.006.
- [23] F. Wen, G. Zhang, L. Sun, X. Wang, and X. Xu, “A hybrid temporal association rules mining method for traffic congestion prediction,” *Comput. Ind. Eng.*, vol. 130, no. 6, pp. 779–787, 2019, doi: 10.1016/j.cie.2019.03.020.
- [24] F. Li, J. Gong, Y. Liang, and J. Zhou, “Real-time congestion prediction for urban arterials using adaptive data-driven methods,” *Multimed. Tools Appl.*, vol. 75, no. 24, pp. 17573–17592, 2016, doi: 10.1007/s11042-016-3474-3.
- [25] J. Song, C. Zhao, S. Zhong, T. A. S. Nielsen, and A. V. Prishchepov, “Mapping spatio-temporal patterns and detecting the factors of traffic congestion with multi-source data fusion and mining techniques,” *Comput. Environ. Urban Syst.*, vol. 77, no. July, 2019, doi: 10.1016/j.compenvurbsys.2019.101364.
- [26] A. A. Heidari, S. Mirjalili, H. Faris, I. Aljarah, M. Mafarja, and H. Chen, “Harris hawks optimization: Algorithm and applications,” *Futur. Gener. Comput. Syst.*, vol. 97, no. March, pp. 849–872, 2019, doi: 10.1016/j.future.2019.02.028.
- [27] R. Venkata Rao, “Jaya: A simple and new optimization algorithm for solving constrained and unconstrained optimization problems,” *Int. J. Ind. Eng. Comput.*, vol. 7, no. 1, pp. 19–34, 2016, doi: 10.5267/ij.ijiec.2015.8.004.
- [28] N. A. M. Yusof et al., “Deep convolution neural network for crack detection on asphalt pavement,” *J. Phys. Conf. Ser.*, vol. 1349, no. 1, 2019, doi: 10.1088/1742-6596/1349/1/012020.
- [29] S. Mirjalili, S. M. Mirjalili, and A. Lewis, “Grey Wolf Optimizer,” *Adv. Eng. Softw.*, vol. 69, pp. 46–61, 2014, doi: 10.1016/j.advengsoft.2013.12.007.
- [30] B. M. Nguyen, T. Tran, T. Nguyen, and G. Nguyen, “Hybridization of Galactic Swarm and Evolution Whale Optimization for Global Search Problem,” *IEEE Access*, vol. 8, pp. 74991–75010, 2020, doi: 10.1109/ACCESS.2020.2988717.
- [31] X. Tang, “Large-scale computing systems workload prediction using parallel improved LSTM neural network,” *IEEE Access*, vol. 7, pp. 40525–40533, 2019, doi: 10.1109/ACCESS.2019.2905634.
- [32] R. Toncharoen and M. Piantanakulchai, “Traffic State Prediction Using Convolutional Neural Network,” *Proceeding 2018 15th Int. Jt. Conf. Comput. Sci. Softw. Eng. JCSSE 2018*, pp. 1–6, 2018, doi: 10.1109/JCSSE.2018.8457359.
- [33] M. Jaihuni et al., “A novel recurrent neural network approach in forecasting short term solar irradiance,” *ISA Trans.*, vol. 121, no. xxxx, pp. 63–74, 2022, doi: 10.1016/j.isatra.2021.03.043.
- [34] B. Karthika, N. Umamaheswari, and R. Venkatesh, “A research of traffic prediction using deep learning techniques,” *Int. J. Innov. Technol. Explor. Eng.*, vol. 8, no. 9 Special Issue 2, pp. 725–728, 2019, doi: 10.35940/ijitee.II151.0789S219.

Impact of the Pandemic on the Development and Regulation of Electronic Commerce in Russia

Svetlana Panasenکو, Maisa Seifullaeva, Ibragim Ramazanov
Elena Mayorova, Alexander Nikishin, A.M. Vovk
Plekhanov Russian University of Economics
Moscow, Russia

Abstract—The article deals with topical issues of legal support for the development of electronic commerce in the Russian Federation. The analysis of the main categories of e-commerce has been carried out, the content of which is standardized in domestic regulatory legal acts, the following is among them: elements of the purchase and sale process, the concept of types of trading activities, electronic signature, digital assets, digital currency, smart contracts, digital transactions, etc.). The categories have been defined, the concept of which is absent in the normative legal acts of Russia: the concept of digital goods, e-commerce infrastructure, e-commerce services, delivery channels in online stores, the concept of a courier/courier service, the concept of smart applications, the definition of varieties of online stores, etc. The problem of research is defined: insufficiently effective legal support of economic activity in electronic commerce. An imperfect system of planning strategies for the development of trade organizations in the online environment has been revealed. The conclusions have been formed that the process of digitalization and the consequences of the pandemic have a significant impact on the dynamics of legal support for the development of electronic commerce, but its level is currently not high enough and requires improvement (by making additions to several regulatory legal acts, such as the Law on Trade, the Strategy for the Development of Electronic Commerce in Russia, etc.). The study used system, situational, complex methods, graphical, block grouping, methods of comparative analysis of normative legal acts, and synthesis of conclusions and proposals.

Keywords—Legal support; strategy; development; electronic commerce; Russian federation; digitalization; informatization; online stores; delivery channels; digital product

I. INTRODUCTION

An important place in the progressive and dynamic development of electronic commerce in Russia is occupied by the issues of optimizing the legislative and regulatory framework that regulates the development of this segment of the country's economy. It is very relevant to determine the level of legal support for the development of electronic commerce in the Russian Federation, identify problems and features of existing regulatory legal acts, and form proposals for their improvement, which will ensure faster progress of electronic commerce, which is the driver of the development of the Russian economy, especially in the era of its digitalization. These issues are particularly relevant given the experience of the pandemic and the way out of it, which created additional impulses for the development of electronic commerce and caused the need for further improvement of information and legal support for this process.

A review of sources on this topic showed that the issues of legal support for electronic commerce are being studied both in Russia and abroad. Foreign authors mainly devote their attention to the peculiarities of the legal regulation of electronic commerce, taking into account the experience of specific foreign countries. For example, Gulomrizoi [1] investigated the legal costs of implementing the law on electronic commerce in the Islamic Republic of Iran. Belyavskaya [2] analyzed the issues of regulation of e-commerce in Ukraine. Mukhopadhy [3] studied the issues of e-commerce and data localization from the perspective of developing countries. Some other foreign authors, such as Olifirova and Yagmur [4] analyzed more narrow issues of legal support for electronic commerce (including cost accounting and calculations in interactive commerce: their organizations and methods). Khasenova [5] studied regional aspects (legislative aspects of the strategic development of the economy of the regions of Kazakhstan).

Russian authors Kandybko, Denisov, Petrochenko [6] focused their attention on the history of the emergence and development of electronic commerce in Russia in the field of public procurement for defense, and, accordingly, investigated the issues of assessing the current state and problems of legal regulation of electronic procurement methods for defense, indicating the positive and negative aspects of the transition to electronic closed bidding procedures using the functionality of specialized electronic platforms and the formation of further ways to improve the legal support of procurement in electronic form (purposefully in the segment of public procurement for the defense of Russia). Other Russian authors, Makarevich and Rudneva [7] focused their research attention on the prospects of using blockchain technology for the development of electronic commerce and the features of its legal support. Most often, researchers studied the features of the legal regulation of electronic commerce for certain types of goods. For example, Shvedov [8] studied the problems of legal security of electronic jewelry trade in the Russian Federation. Domestic authors paid attention to certain aspects of electronic commerce in other works, for example, procurement activities in the study of Andreeva [9], the use of innovative mechanisms in electronic commerce, and the regulation of these issues [10], legal regulation of the use of intangible assets in electronic commerce [11], regional features of the development of electronic commerce and consideration these features in local regulatory legal acts [12], the peculiarities of rationing of the retail trade network [13].

It should be noted that there are works by various authors on the study of the legal support of electronic commerce in Russia as a whole [14, 15, 16, 17], but such works are not enough taking into account the latest and not fully researched the experience of the pandemic. Therefore, there is a need to conduct such studies, which would allow a more balanced approach to the issues of modern legal support of electronic commerce in Russia in the future in the process of gradual recovery from the pandemic.

The following research problem has been identified: insufficiently effective legal support for economic activity in electronic commerce, as well as an imperfect system for planning strategies for the development of trade organizations in the online environment. This problem has significantly worsened against the background of the manifestation of the crisis phenomena associated with the pandemic.

The research objective is to identify the features of legal support for the development of electronic commerce in Russia and identify areas for its improvement, taking into account the experience of the pandemic. Research objectives: the systematization of the experience of domestic and foreign authors on the issues of legal support for the development of electronic commerce, conducting a comparative analysis of the fundamental legislative acts regulating the legal support for the development of electronic commerce in Russia, forming conclusions and proposals for improving the legal support for the development of electronic commerce in Russia, taking into account the experience of the pandemic. The research object: electronic commerce. Research subject: legal support for the development of electronic commerce in Russia.

The article describes the results of the study: analysis of the legislative consolidation of key elements of e-commerce processes, analysis of strategic documents on the development of electronic commerce, analysis of the legal regulation of e-commerce categories related to remote biometric identification and signature, conclusion with conclusions and recommendations.

II. METHODS

The study used system, situational, complex methods, graphical, block grouping, methods of comparative analysis of normative legal acts, and synthesis of conclusions and proposals. The system method was used to analyze all elements of the system of legal support for electronic commerce in the Russian Federation, the situational approach was used to determine the current impact of the pandemic on the development of trading activities in the online environment, the graphical method was used to create an effective visualization of the results of the study, the block grouping method allowed us to identify groups of concepts of electronic commerce that are meaningfully enshrined in the current regulations in Russian legislation and groups of concepts, which are absent (while trade practice requires their further legislative consolidation), the method of comparative analysis allowed for a consistent analysis of various legislative and regulatory acts on electronic commerce (especially in terms of the rules for conducting trade transactions in an online environment).

III. RESULT AND DISCUSSION

The study of the main categories of the e-commerce process in the modern legislative and legal field of Russia in the context of the transition to the digital economy has shown that the digitalization process and the consequences of the pandemic have a significant impact on their dynamics.

A. Analysis of the Legislative Consolidation of Key Elements of E-Commerce Processes

It has been revealed that the key categories of elements of e-commerce processes, such as purchase and sale, informatization, and others, are legally fixed. The analysis showed that the Civil Code of the Russian Federation [18] describes various rights (ownership, use, disposal of property). In addition, the concept of payment for goods is noted, the sale of goods based on the remote method of selling goods, the buyer's rights in the event of the sale of goods of improper quality to him/her, the concept of providing information about the goods to the buyer is interpreted.

The Arbitration Procedure Code of the Russian Federation [19] interprets the right to appeal to the arbitration court to protect their rights. Economic disputes (which in practice are addressed by trade organizations, including in the field of commodity circulation) can also be resolved in a pre-trial order. The arbitration court allows considering the documents signed with an electronic signature.

Federal Law No. 381-FZ in Article 2 contains several concepts that are relevant to the structural elements of the competitive environment of the sphere of commodity circulation (Table I) [20].

In our opinion, the absence of terms in Article 2 directly related to various types of trade as part of the sphere of commodity circulation (including electronic commerce) is among the shortcomings of Federal Law No. 381-FZ [20]. This Federal Law does not contain the concepts of e-commerce, online store, marketplace, etc.

Federal Law No. 487-FZ and Articles 5 and 8 of Federal Law No. 381 contain the concept of the product code [20, 22]. The same Federal Law indicates the list of goods subject to labeling.

TABLE I. THE CONCEPTS SPECIFIED IN FEDERAL LAW NO. 381 AND GOST R 51303-2013

The concepts specified in Federal Law N 381 [20]	The concepts specified in GOST R 51303-2013 [21]
Trading activity, wholesale trade, retail trade, retail network, food products, goods marked with identification means, goods subject to mandatory marking by means of identification, identification tool, marking code, identification code, verification code	Wholesale and retail trade, consumer market, buyer, wholesale buyer, seller, trade organization, trading company with a specialized assortment, trading company with a combined assortment, trading company with a mixed assortment of goods, distribution center warehouse, general goods warehouse, specialized warehouse, universal warehouse

Article 2 in Federal Law No. 149-FZ presents the following basic concepts that are relevant to electronic commerce: a website on the Internet, a domain name, a network address, etc. [23].

Federal Law No. 63-FZ (the latest version) uses the following basic concepts that are important for electronic commerce: electronic signature, electronic signature means, participants in electronic interaction, etc. [24].

Federal Law No. 2300-1 includes a description of the following terms that are also used in electronic commerce: consumer, manufacturer, seller, lack of goods, owner of an information aggregator, etc. [25]. The presence of legislative consolidation of such fundamental concepts is of particular importance in the process of overcoming the pandemic (in the context of the growth of both effectively operating online stores that have an excellent reputation among consumers, the expansion of the activities of aggregators (marketplaces), and the growth of fictitious and unreliable online organizations selling goods of questionable quality).

Further, the study showed that GOST R 51303-2013 "Trade. Terms and definitions" contains a much more complete list of terms related to trade, but only a part of them can be indicated among those related to electronic commerce [21]. The concepts of trading activity, wholesale trade, retail trade, retail chains, food products are interpreted in GOST in the same way as in Federal Law No. 381-FZ. The following concepts are given separately in GOST (Table I).

The GOST indicates the difference in terms of distance selling, mail-order trading, and e-commerce. GOST also gives a clear definition of what an electronic trading procedure is. GOST also indicates the concept of online commerce. The advantages of GOST, in our opinion, should include an extremely wide list of basic trade concepts that are fully applied in electronic commerce (goods, price, assortment, types of assortment, types of prices, quality of goods, etc.). At the same time, a major drawback of the State Standard for Trade is the lack of concepts of various types of online stores (as retail objects). In addition, a significant disadvantage and drawback of these regulatory legal acts is the lack of concepts of digital goods, e-commerce infrastructure, the concept of information about the product presented on official Internet pages (this is especially important for describing and familiarizing with the product on the website of the online store, which excludes tactile and physical contact with the product, possible distortion of colors during electronic color rendering and other restrictions). There is no concept of electronic trading services, online purchases, delivery channels in online stores, the concept of a courier or courier service, electronic means of payment in GOST and Federal Law No. 381-FZ.

B. Analysis of Strategic Documents on the Development of Electronic Commerce

A study of strategic documents on e-commerce in Russia showed that certain shortcomings were eliminated in the draft Strategy for the development of e-commerce until 2025, developed in 2017, which proposed a description of such terms as digital economy, digital economy ecosystem, e-commerce

(B2C e-commerce sector), use of the Internet channel by retail, online store, cross-border e-commerce, online purchase, digital goods, mobile commerce, B2G e-commerce sector, B2B e-commerce sector, electronic trading platform (marketplace), e-commerce transaction, O2O business (from "online to offline"), machine-to-machine interaction (M2M), aggregator of goods (services) [26].

Some terms were excluded in the updated version of the draft Strategy for the development of Electronic Commerce (as part of the Strategy for the Development of Trade in the Russian Federation for 2019-2025, published in September 2019 [27]). It should be noted that in the latest version of the Strategy for the Development of Trade in the Russian Federation for 2019-2025, the terms for e-commerce are not given in the text of the Strategy for the Development of e-commerce, but in the form of an Appendix to the entire Strategy for the Development of Trade and currently includes the following terms: digital economy, electronic commerce, e-commerce, internet trading, online store, cross-border e-commerce, B2G e-commerce sector, wholesale e-commerce (B2B), electronic trading platform (marketplace) [27].

Let us present the terms that were excluded from the latest version of the draft Strategy for the Development of Electronic Commerce until 2025 (published in September 2019): an ecosystem of the digital economy, use of the online retail channel, online shopping, digital goods, mobile commerce, electronic transaction, O2O business (from "online to offline"), machine-to-machine interaction (M2M), aggregator of goods (services). In our opinion, the exclusion of certain terms from the new version of the draft Strategy for the Development of Electronic Commerce until 2025 (published in 2019) (digital product, Internet purchase, use of an Internet channel, electronic transaction) made the last list of terms on electronic commerce not sufficiently complete and meaningful [28].

In addition, it should be noted as a disadvantage that the latest version of the E-Commerce Development Strategy does not justifiably contain a description of such concepts as electronic money, electronic product information, electronic signature, e-commerce infrastructure, various types of online stores, information protection, electronic means of payment (for example, an electronic wallet), digital transactions [27].

Let us complement: the definitions of the Internet channel and Internet purchases are neither reflected in GOST R 51303-2013 "Trade. Terms and definitions" nor in the latest version of the Draft Strategy for the Development of Electronic Commerce in the Russian Federation until 2025. The urgent need to interpret these terms, taking into account the experience of the pandemic, is of high importance.

The definition of the B2C sector in the draft Strategy for the Development of Electronic Commerce until 2025, and in its updated version from 2019 is not quite correctly formulated in terms of informing the seller about the intention to buy a product or give feedback after the purchase (which in practice can occur both online and offline).

In our opinion, it is necessary to distinguish the terms Internet commerce (as a form of electronic commerce) and the B2C Sector (as a special type of sale and/or provision of

services to the end consumer individuals). Currently, these terms are being mixed in the modern version of the draft Strategy for the Development of Electronic Commerce until 2025.

The analysis also showed that the terms "Information, Information Technology and Information Protection" are described neither in GOST R 51303-2013 "Trade. Terms and definitions", nor in the latest version of the Draft Strategy for the Development of Electronic Commerce in the Russian Federation until 2025, nor in the Federal Law of December 28, 2009, No. 381-FZ. It is necessary to focus on Federal Law No. 149-FZ, which presents the following concepts: information, the confidentiality of information, a website on the Internet, a domain name, a network address, the owner of a website on the Internet, a hosting provider, a unified identification and authentication system, a search engine, etc. [23].

In addition, the analysis showed that the "information protection" term in Federal Law No. 149-FZ is not considered in the list of terms in Article 2, but is interpreted in Article 16: information protection is the adoption of legal, organizational, and technical measures.

Additional information in this regard is provided by the Decree of the President of the Russian Federation No. 646 and the Decree of the President of the Russian Federation No. 203 [29, 30]. The following basic concepts are used in the Information Security Doctrine of the Russian Federation: information security, means of ensuring it, etc.

The Decree of the President of the Russian Federation No. 203 includes the following definitions: information society, information space, processing of large amounts of data.

Thus, as the analysis of the terminology for the information block shows in terms of definitions that would refract and take into account the specifics of e-commerce in the digital economy in the face of recovering from the pandemic, this direction is insufficiently provided (although it is information that becomes the most important factor in the development of the entire economy in general, and its segments in particular). This is especially significant in the era of the development of information technologies, knowledge, and intelligence, in the process of overcoming the crisis associated with the pandemic. In our opinion, it is necessary to add a description of terms for the protection of information in electronic commerce (as a practice of preventing unauthorized access, use, disclosure, distortion, modification, research, recording, or destruction of information in electronic transactions for the sale and purchase) and information security in the legislative and legal field of Russia.

The analysis showed that the "E-commerce infrastructure" term is also not sufficiently developed. This term ("Infrastructure of electronic commerce") is legally described neither in GOST R 51303-2013 "Trade. Terms and definitions", nor in the latest version of the Draft Strategy for the Development of Electronic Commerce in the Russian Federation until 2025, nor in Federal Law No. 381-FZ.

In addition, the text of the draft Strategy for the Development of Electronic Commerce until 2025 mentions the technical and logistics infrastructure of electronic commerce in

the list of directions for the development of electronic commerce, but the interpretation of these concepts is not given in the list of terms [27].

In our opinion, to improve the legal support of electronic commerce in Russia, we should propose a structural-block approach to describing the elements of the electronic sales format infrastructure (which should be reflected in the relevant regulatory legal acts of the Russian Federation):

- 1) Technical and technological block (including software and platform software);
- 2) The communication unit (for electronic exchange and use of data, including in promotion);
- 3) Logistics unit (aspects of warehousing and delivery);
- 4) Electronic transaction block;
- 5) The HR support unit with the predominance of digital competencies of personnel.

The combination of these elements makes up the infrastructure of the electronic sales format in the field of circulation in the digital economy. These definitions are recommended for inclusion in the list of terms, as it is one of the key ones that allows presenting a complex of the most important structural components of electronic sales in electronic commerce in the digital economy in the long-term (strategic) perspective in the process of overcoming the pandemic.

C. Analysis of the Legal Regulation of E-Commerce

Categories Related to Remote Biometric Identification and Signature

The study showed that categories such as "Remote biometric identification and cloud signature", which have become urgently in demand in the pandemic and in the process of recovering from it, are relatively new terms, which in practice require high technologies of the latest generation.

The remote identification mechanism was developed by the Bank of Russia as part of the implementation of the main directions of the development of financial technologies [31]. The creation and development of a platform for remote identification make it possible to transfer financial services to a digital environment, increase the availability of financial services for consumers, including people with disabilities, the elderly, and the disabled, as well as increasing competition in the financial market.

Biometric data of a citizen should be stored in an impersonal form, and separately from personal data, which significantly increases the level of security. User data is transmitted via secure communication channels and placed in the almost impenetrable cloud infrastructure of the system operators.

It should be noted that concerning the terminology of cloud technologies, a significant contribution was made by the Decree of the President of the Russian Federation No. 203, which includes definitions of cloud computing, processing large amounts of data, etc. [30].

An analysis of another category of e-commerce – "cloud signature" – showed that this concept is also quite new and

rapidly developing in the conditions of recovering from the pandemic. A cloud signature is an analog of an electronic signature that has all its properties and functions, but with one significant difference – the cloud signature certificate is stored not on a token or a smart card, but the server of the certification center. Accordingly, the document signing process also takes place remotely, the user only needs to confirm the operation using a mobile application or entering a one-time password on any device connected to the Internet. Cloud signing does not require additional devices and software tools (tokens, password generators, electronic signature software, scratch cards).

A cloud signature has several advantages over an electronic signature:

1) The cloud electronic signature is not linked to a specific computer. The signature is placed on the server of the certification center in a certified secure cell ("cloud") provided to the client for use. It can be accessed from any device (for example, from an iOS or Android smartphone).

2) Cloud-based encryption technologies minimize the cost of buying tokens, as well as the cost of installing and updating special cryptographic software necessary for servicing an electronic signature certificate.

3) Increased reliability and protection due to the refusal to use the key carrier. The cloud signature cannot be lost or broken, and the security is fully provided by the certification center.

In other words, cloud signature is the optimal solution that allows securely signing documents, giving them legal significance, at any time and from any device.

With the help of a cloud signature system in the field of circulation in the digital economy, customers' intentions to submit applications for participation in procurement procedures, authentication, creation and execution of documents, the facts of receiving and/or familiarizing users with certain information can be confirmed.

A Certificate of compliance was issued by the FSB of Russia for performing actions with authentication and confirmation of cloud signature generation operations using mobile applications for IOS and Android.

The experience of using remote biometric identification and cloud signature is in its infancy on the territory of Russia, but in the future, the process of overcoming the pandemic has the potential to grow, including in electronic commerce in the digital economy, so these terms are also recommended for inclusion in the list of necessary terms and categories in the future.

In addition, the analysis of such categories of electronic commerce in the digital economy as: "Smart contracts, smart applications" showed that their emergence and use is associated with the development of "smart" technologies. Federal Law (draft) No. 419059-7 "On Digital Financial Assets" [32] was prepared to regulate this activity in Russia, which proposes the following basic concepts: smart contract, digital wallet, digital financial asset, cryptocurrency, etc.

There is no concept of smart applications in the draft Federal Law No. 419059-7 "On Digital Financial Assets", but there is an urgent need for its development, especially taking into account the experience of the pandemic. So far, there are separate ideas about the essence of this definition. It has not fully developed yet. We can offer the following as one of the interpretations: smart applications are software specially developed for a specific platform (iOS, Android, Windows Phone, etc.), which is intended for use on smartphones, tablets, and other devices and provides unique opportunities for expanding the functionality of mobile devices.

Federal Law No. 259-FZ was prepared based on Federal Bill No. 419059-7 [33]. Taking into account the rapid growth of the practice of using smart technologies, such categories as digital assets, digital currency, smart contracts will gradually enter the field of e-commerce in the digital economy, taking into account the experience of the pandemic.

It should also be noted that an analysis of another category, such as "Public networks", was performed, which showed that one should be guided by Federal Law No. 126-FZ [34]. According to the law, the Internet is a general-purpose network. In addition, the Decree of the President of the Russian Federation No. 203 gives the following definition to the varieties of this category: industrial Internet, new generation communication networks, the Internet of Things [30].

Thus, the analysis of the entire complex of categories relevant to online sales showed that these concepts and their content are not fully covered by the regulatory legal acts of Russia regulating the development of electronic commerce in the Russian Federation.

IV. CONCLUSION

We conclude that the considered categories of elements of the e-commerce process in the Russian Federation in the digital economy are highly significant and relevant, but the level of legal support for the development of e-commerce is not at a high level and requires several improvements.

On the one hand, the categories of e-commerce processes were identified, the content of which is normalized in regulatory legal acts (sales of goods, their payment, buyer's rights, the concept of electronic signature, digital assets, digital currency, smart contracts, digital transactions, etc.). At the same time, it was revealed that several categories related to electronic commerce are absent in the regulatory legal acts of Russia: there is no concept of digital goods, e-commerce infrastructure, e-commerce services, delivery channels in online stores, the concept of a courier/courier service, the concept of smart applications, the definition of varieties of online stores, etc.

As a result, the conclusions have been formed: even though the process of digitalization and the consequences of the pandemic have a significant impact on the dynamics of legal support for the development of electronic commerce, its level in Russia is not yet high enough and requires improvement. For example, it requires the revision of the GOST on trade (which should include such concepts of various types of online stores, digital goods, e-commerce infrastructure, the concept of product information presented on official Internet pages, the

concept of electronic trading services, online purchases, delivery channels in online stores, the concept of a courier or courier service, electronic means of payment. Strategic documents on the development of e-commerce also need to be finalized: it would be desirable to include a description of the following concepts in the e-Commerce Development Strategy: electronic money, electronic product information, electronic signature, e-commerce infrastructure, various types of online stores, information protection, electronic means of payment (for example, an electronic wallet), digital transactions. The Law on Trade in the Russian Federation also needs additions (which should include such terms as the concept of e-commerce, online store, marketplace, etc.).

Timely revision and improvement of the conceptual and categorical apparatus in the legal support of electronic commerce in Russia, its unification, consolidation of these concepts and their content in the Russian regulatory framework, their wider use in domestic post-pandemic practice will eliminate objective obstacles to faster development of electronic commerce in the Russian economy, especially taking into account the pandemic-related crisis recovery and the acceleration of digitalization processes.

Among the prospects of our research, it should be noted the need to continue scientific and analytical work in the direction of studying the changes being made to the legal support of electronic commerce in the conditions of exiting the pandemic at the present time, in the medium and long term (taking into account the prolonged cumulative impact of the pandemic on the development of electronic commerce). In the future, there is also the study of changing strategic documents on e-commerce at the national level, taking into account the process of digitalization, the study of various strategic alternatives for the development of e-commerce, its strategic trajectories. Promising, in our opinion, is not only the continuation of the study of the All-Russian level of development, but also regional (the level of individual large regional entities - subjects of the Russian Federation). This will make the study more comprehensive, in-depth and meaningful in the future.

ACKNOWLEDGMENT

The research was carried out within the framework of the state task of the Ministry of Science and Higher Education of the Russian Federation FSSW-2020-0009 "Development of a methodology for managing the competitiveness of enterprises in the field of commodity circulation in the digital economy".

REFERENCES

- [1] G. Gulomrizoi, "Pravovye izderzhki ispolneniya zakona ob elektronnoi trgovle v islamskoi respublike Iran [Legal costs of enforcing the law on electronic commerce in the Islamic Republic of Iran]," *Vestnik Tadzhijskogo natsionalnogo universiteta*, no. 3-6, pp. 97-99, 2013.
- [2] Yu. V. Belyavskaya, "Regulation of e-commerce in Ukraine [Regulirovanie elektronnoi kommertsii v Ukraine]," *Molodii vchenii*, no. 10(37), pp. 336-339, 2016.
- [3] A. Mukhopadhyaya, "E-commerce and data localization: The position of developing countries [Elektronnaya trgovlya i lokalizatsiya dannykh: Pozitsiya razvivayushchikhsya stran]," *Vestnik mezhdunarodnykh organizatsii: Obrazovanie, nauka, novaya ekonomika*, vol. 15, no. 3, pp. 153-175, 2020.
- [4] Yu. A. Olifirova, and K. A. Yagmur, "Accounting for costs and settlements in online trading: Organization and methodology [Uchet zatrat i raschetov v interaktivnoi trgovle: Organizatsiya i metodika],"

Bulletin of the Mykhailo Tuhan-Baranovsky Donetsk National University of Economics and Trade, no. 3(59), pp. 154-164, 2013.

- [5] K. E. Khasenova, "Legislative aspects of the strategic development of the regions of Kazakhstan [Zakonodatelnye aspekty strategicheskogo razvitiya regionov Kazakhstana]," in *Proceedings of the Sixth International Scientific and Practical Conference "Problems and Prospects for the Development of Economics and Management in Russia and Abroad" [Sbornik trudov Shestoi mezhdunarodnoi nauchno-prakticheskoi konferentsii "Problemy i perspektivy razvitiya ekonomiki i menedzhmenta v Rossii i za rubezhom"]*, O. P. Osadchaya, Ye. S. Belyayeva, and D. V. Remizov, Eds. Rubtsovsk: Rubtsovsk Industrial Institute, 2014, pp. 301-311.
- [6] N. V. Kandybko, D. B. Denisov, and A. I. Petrosenko, "Legal aspects of the use of electronic commerce technologies in public procurement to ensure defense [Pravovye aspekty primeneniya tekhnologii elektronnoi trgovli v gosudarstvennykh zakupkakh dlya obespecheniya oborony]," *Vestnik voennogo prava*, no. 1, pp. 51-58, 2021.
- [7] M. L. Makarevich, and T. S. Rudneva, "Prospects for the use of blockchain technology for the development of electronic commerce and features of its legal support [Perspektivy ispolzovaniya tekhnologii blokchein dlya razvitiya elektronnoi trgovli i osobennosti ee pravovogo obespecheniya]," in *SPbPU Science Week. Materials of a scientific conference with international participation. Institute of Industrial Management, Economics and Trade [Nedelya nauki SPbPU. Materialy nauchnoi konferentsii s mezhdunarodnym uchastiem. Institut promyshlennogo menedzhmenta, ekonomiki i trgovli]*, St. Petersburg: St. Petersburg Polytechnic University of Peter the Great, 2018, pp. 238-241.
- [8] V. V. Shvedov, "Problems of legal support for the security of electronic trade in jewelry in the Russian Federation [Problemy pravovogo obespecheniya bezopasnosti elektronnoi trgovli yuvelirnymi izdeliyami v Rossiiskoi Federatsii]," in *Economic and legal problems of ensuring economic security. Materials of the All-Russian Scientific and Practical Conference [Ekonomiko-pravovye problemy obespecheniya ekonomicheskoi bezopasnosti. Materialy Vserossiiskoi nauchno-prakticheskoi konferentsii]*, E. G. Animitsa, and G. Z. Mansurov, Eds. Yekaterinburg: Ural State University of Economics, 2018, pp. 220-223.
- [9] L. V. Andreeva, "Elements of digital technologies in trade and procurement (legal aspect) [Elementy tsifrovyykh tekhnologii v trgovoi i zakupochnoi deyatel'nosti (pravovoi aspekt)]," *Prilozhenie k zhurnalu Predprinimatelskoe pravo*, no. 1, 15-21, 2019.
- [10] S. V. Panasenko, V. P. Cheglov, I. A. Ramazanov, E. A. Krasilnikova, I. B. Stukalova, and A. V. Shelygov, "Improving the innovative development mechanism of the trade sector," *Journal of Advanced Pharmacy Education and Research*, vol. 11, no. 1, pp. 141-146, 2021.
- [11] S. Panasenko, O. Karashchuk, E. Krasilnikova, E. Mayorova, and A. Nikishin, "Analysis of intangible assets of online stores in Russia," *International Journal of Management*, vol. 11, no. 5, pp. 579-589, 2020.
- [12] S. V. Panasenko, O. S. Karashchuk, E. A. Mayorova, A. F. Nikishin, and A. V. Boykova, "Regional aspects of e-commerce development," *International Journal of Civil Engineering and Technology*, vol. 10, no. 2, pp. 1821-1829, 2019.
- [13] O. Karashchuk, I. Nusratullin, V. Tretyakov, M. Shmatov, and A. Rezvan, "Retail chains in Russia: Some aspects of state regulation," *Journal of Advanced Research in Law and Economics*, vol. 10, no. 4, pp. 1258-1265, 2019. [https://doi.org/10.14505/jarle.v10.4\(42\).25](https://doi.org/10.14505/jarle.v10.4(42).25).
- [14] N. V. Alekseeva, N. B. Andrenov, V. P. Beklemishev, N. V. Butayeva, A. V. Vavilina, M. V. Ginsburg, et al., "Economic research: Analysis of the state and development prospects [Ekonomicheskie issledovaniya: Analiz sostoyaniya i perspektivy razvitiya]": A monograph, Vol. Book 47. Voronezh: Voronezh State Pedagogical University; Moscow: Nauka: Inform, 2018.
- [15] L. A. Bragin, G. G. Ivanov, S. V. Panasenko, L. A. Efimovskaya, O. S. Karashchuk, E. A. Krasilnikova, et al., *Globalization of trade based on innovations: A monograph*. Hamilton: Accent Graphics Communications & Publishing, 2018.
- [16] S. V. Panasenko, and T. A. Mazunina, "Trends in the development of modern trade [Tendentsii razvitiya sovremennoi trgovli]," in *Modern trade: Theory, practice, innovation. Materials of the 7th All-Russian Scientific and Practical Conference with International Participation*

- [Sovremennaya trgovlya: Teoriya, praktika, innovatsii. Materialy 7 Vserossiiskoi nauchno-prakticheskoi konferentsii s mezhdunarodnym uchastiem], E. V. Gordeeva, S. V. Porosenkov, and V. N. Yakovlev, Eds. Perm: Perm Institute (branch) of the Russian Economic University named after G.V. Plekhanov, 2017, pp. 62-68.
- [17] M. E. Seifullaeva, S. V. Panasenko, I. P. Shirochenskaya, A. B. Tsvetkova, and J. Yevseyeva, "Main tendencies and problems of agricultural export and import in Russia under economic sanctions," *Espacios*, vol., 39, no. 9, p. 38, 2018.
- [18] State Duma of the Federal Assembly of the Russian Federation, The Civil Code of the Russian Federation of November 30, 1994 No. 51-FZ. *Sobranie Zakonodatel'stva Rossiiskoi Federatsii [SZ RF] [Collection of Legislation of the RF] No. 32, Item 3301, 05.12.1994.*
- [19] State Duma of the Federal Assembly of the Russian Federation, Arbitration Procedural Code of the Russian Federation of July 24, 2002 No. 95-FZ. *Sobranie Zakonodatel'stva Rossiiskoi Federatsii [SZ RF] [Collection of Legislation of the RF] No. 30, Item 3012, 29.07.2002.*
- [20] State Duma of the Federal Assembly of the Russian Federation, Federal Law of December 28, 2009 No. 381-FZ (as amended on December 25, 2018) "On the basics of state regulation of trading activities in the Russian Federation". *Sobranie Zakonodatel'stva Rossiiskoi Federatsii [SZ RF] [Collection of Legislation of the RF] No. 1, Item 2, 04.01.2010.*
- [21] Federal Agency for Technical Regulation and Metrology, GOST R 51303-2013 Trade. Terms and definitions. Moscow: Standartinform, 2013.
- [22] State Duma of the Federal Assembly of the Russian Federation, Federal Law of December 31, 2017, No. 487-FZ "On amendments to Article 47 of the Federal Law "On the use of cash register equipment in cash settlements and (or) settlements using electronic means of payment". *Rossiiskaia Gazeta [Ros. Gaz.] No. 1(7464), 09.01.2018.*
- [23] State Duma of the Federal Assembly of the Russian Federation, Federal Law of July 27, 2006 No. 149-FZ (as amended on May 1, 2019) "On information, information technologies and information protection". *Sobranie Zakonodatel'stva Rossiiskoi Federatsii [SZ RF] [Collection of Legislation of the RF] No. 31 (Part 1), Item 3448, 31.07.2006.*
- [24] State Duma of the Federal Assembly of the Russian Federation, Federal Law of April 6, 2011, No. 63-FZ "On electronic signature". *Sobranie Zakonodatel'stva Rossiiskoi Federatsii [SZ RF] [Collection of Legislation of the RF] No. 15, Item 2036, 11.04.2011.*
- [25] Supreme Soviet of the Russian Federation, Federal Law of February 7, 1992 No. 2300-1 "On Consumer Rights Protection". *Sobranie Zakonodatel'stva Rossiiskoi Federatsii [SZ RF] [Collection of Legislation of the RF] No. 3, Item 140, 15.01.1996.*
- [26] Ministry of Industry and Trade of Russia, Draft Strategy for the development of electronic commerce until 2025. 2017. [Online]. Available: https://minpromtorg.gov.ru/docs/#!proekt_strategiya_razvitiya_elektronnoy_torgovli_v_rossiyskoy_federatsii_na_period_do_2025_goda [Accessed: August 8, 2021].
- [27] Ministry of Industry and Trade of Russia, Draft Strategy for the development of trade in the Russian Federation until 2025. 2019. [Online]. Available: https://minpromtorg.gov.ru/press-centre/news/!opublikovan_proekt_strategii_razvitiya_torgovli_do_2025_goda [Accessed: August 8, 2021].
- [28] V. P. Cheglov, S. V. Panasenko, E. A. Krasilnikova, E. A. Mayorova, and S. Yu. Kazantseva, "Modern trends in the development of markets of goods and services No. 7," Article in the open archive, 7, p. 13, 2020.
- [29] President of the Russian Federation, Decree of the President of the Russian Federation of December 5, 2016 No. 646 "On the approval of the Information Security Doctrine of the Russian Federation". 2016. [Online]. Available: <http://publication.pravo.gov.ru/Document/View/0001201612060002> [Accessed: August 8, 2021].
- [30] President of the Russian Federation, Decree of the President of the Russian Federation of May 9, 2017 No. 203 "On the Strategy for the development of the information society in the Russian Federation for 2017-2030". 2017. [Online]. Available: <http://publication.pravo.gov.ru/Document/View/0001201705100002> [Accessed: August 8, 2021].
- [31] Central Bank of the Russian Federation, Remote identification. 2021. [Online]. Available: https://cbr.ru/fintech/digital_biometric_id/ [Accessed: August 8, 2021].
- [32] Federal Law (draft) of March 20, 2018 No. 419059-7 "On digital financial assets". 2018. [Online]. Available: <https://sozd.duma.gov.ru/bill/419059-7> [Accessed: August 8, 2021].
- [33] State Duma of the Federal Assembly of the Russian Federation, Federal Law of July 31, 2020 No. 259-FZ "On digital financial assets, digital currency and on Amendments to Certain Legislative Acts of the Russian Federation". *Sobranie Zakonodatel'stva Rossiiskoi Federatsii [SZ RF] [Collection of Legislation of the RF] No. 31 (Part 1), Item 5018, 03.08.2020.*
- [34] State Duma of the Federal Assembly of the Russian Federation, Federal Law of July 7, 2003 No. 126-FZ (as amended on June 6, 2019) "On communications". *Sobranie Zakonodatel'stva Rossiiskoi Federatsii [SZ RF] [Collection of Legislation of the RF] No. 28, Item 2895, 14.07.2003.*

Application for a Waste Management via the QR-Code System

Pichit Wandee¹, Zakon Bussabong², Seksit Duangkum³

Department of Information Technology, Faculty of Science Buriram Rajabhat University (BRU), Buriram Province, Thailand¹

Department of Computer Science, Faculty of Science, Buriram Rajabhat University (BRU), Buriram Province, Thailand²

Department of Public Health, Faculty of Science, Buriram Rajabhat University (BRU), Buriram Province, Thailand³

Abstract—This research aims in developing an application for the waste management via the QR code system: 1) to study the quality of the application and 2) to study the satisfaction of users of the application by using the system development life cycle (SDLC) principle. There were 388 people of sample groups in this research which consisted of community leaders, village health volunteers, youth and the general public of Ban Yang Sub-district, Mueang Buriram District, and Buriram Province. The research instruments were the application for a waste management via the QR code system, Application Quality Assessment Form, and the application satisfaction questionnaire. The statistics used in the data analysis were mean and standard deviation. The results of the research revealed that there were three main functions of the application for a waste management via the QR code system as follows: 1) The quality assessment of the application in all aspects at a high level (Mean = 4.41, S.D.=0.10). 2) Study of satisfaction of the application users in all aspects at a high level (Mean = 4.42, S.D. = 0.45). 3) The application of waste management application via the QR code system allowed group members in the community to reduce the process of managing household waste more conveniently and create a positive attitude in using waste to elevate through the work of the group members in the community.

Keywords—Application; QR-Code system; waste management; SDLC; community information

I. INTRODUCTION

Ban Yang Subdistrict Administrative Organization, Mueang Buriram District, Buriram Province has an area of approximately 30.24 square kilometers and a total population of 13,397 [1]. The community in Ban Yang is a model for waste management by sorting and has a management process within the group with the support of the Ban Yang Subdistrict Administrative Organization in which Panrak Pansuk Waste Bank was established. Household waste is classified as wet waste, general waste, recyclable waste and hazardous waste. This causes reduced waste in Ban Yang Sub-district and can be managed within the group using rules and agreements which has clearly divided the structure of the group, such as general administrative subdivision, sales management, and accounting and finance. Source operating profit rewards were shared among members within the group and used to manage other aspects of community welfare, such as funerals, scholarships, care for the poor, the elderly and the disabled, etc. From the strong operation of the group resulted in various awards, such as the Dhamma and Golden Land Village Award 1992, the Outstanding Development and Environmental Conservation

Village Award and the Outstanding Community Plan Village 2008 [2].

The development and expansion of the community since 1980 has resulted in the amount of waste, leftovers, scrap materials, as well as dense living; therefore, increasing population. This results in waste within the community of 1.52 tons per month, causing the cost of transportation in order to get rid of the waste approximately 277,120 Baht per year. This in turn causes environmental pollution problems and the problem of dengue fever due to the large number of Aedes mosquitoes in the community. In the past, the community had managed the waste by segregating it among households. However, it was always difficult to find a place to dispose of the waste. The municipality of Ban Yang Subdistrict, therefore had to dispose of it in the landfill of the municipality of Satuk Subdistrict, which is far away. This results in higher management costs. Furthermore, the behavior of community members remains unaware of the need to maintain a sanitary waste disposal system. The resolution of problems in the past had to depend mainly on government agencies, also the lack of integration of work, and the involvement of all agencies in concrete [3]. According to group leaders on waste management, neighboring communities still lack adequate waste management practices. As a result, the amount of waste in the community continues to increase.

The software development life cycle System Development Life Cycle (SDLC) means development of the system has been set to go in the same direction and set a procedure that is a guideline for analyzing the system by trying to have as few flaws as possible because the current system analysis work is more complicated than in the past. System analysts have to standardize in the development of such work system. Therefore, a system development cycle has been invented to meet the needs of system analysts. SDLC consists of 1) problem definition, 2) analysis, 3) design, 4) development, 5) testing, 6) implementation, and 7) maintenance. This provides an effective framework and method to develop software applications. This helps in effective planning before starting the actual development. SDLC allows developers to analyze the requirements, helps in reducing unnecessary costs during development. During the initial phases, developers can estimate the costs and predict mistakes which may cause expenses, enables developers to design and build high-quality software products. This is because they follow a systematic process that allows them to test the software before it is rolled

out, provides a basis when evaluating the effectiveness of the software. This further enhances the software product [4].

II. LITERATURE REVIEW

Developing the application for a waste management via the QR code system proposed the concept, related theories and research as follows:

1) *Application*: Application is a program that facilitates various aspects designed for Mobile phones, tablets, or any other mobile communication device that we know [5] – [8]. In each operating system, there will be many application developers to meet the needs of users. The author in [9] is available for both free and paid downloads including education, communication, or even entertainment, etc. Mobile applications are divided into three types which are Native Application, Hybrid Application and Web Application.

- Native Application is an application developed with a Library or SDK [10], an instrument for developing programs or applications of the operating system on a mobile phone, OS Mobile (Operating System Mobile), especially, for example, Android uses Android SDK, iOS uses Objective C and Windows Phone uses C#, etc. [11] – [17].
- Hybrid Application [18] – [19] is an application which has been developed for the purpose in order to run on every operating system by using the framework to be able to work on all operating systems [20] – [21].
- Web Application is an application [22] that is written as a browser for the use of various web pages, which are customized to display only the necessary parts in order to reduce the processing resources of the device of a smartphone or tablet. This enables the website load faster. In addition, users can use the internet and intranet in low speed as well [23] – [27].

2) *Theory of System Development Life Cycle: SDLC* [28]

- Planning: The purposes of this phase are to find out the scope of the problem and determine solutions, resources, costs, time, benefits and other items which should be considered here [29].
- System Analysis and Requirements: The second phase is where teams consider the functional requirements of the project of the solution [30].
- System Design: The third phase describes, in detail, the necessary specifications, features and operations that will satisfy the functional requirements of the proposed system which will be in place [31].
- Development [32]: Now the real work begins. The development phase marks the end of the initial section of the process. Additionally, this phase signifies the start of the production. The development stage is also characterized by instillation and change.
- Integration & Testing [33]: This phase involves systems integration and system testing (of programs

and procedures) normally carried out by a Quality Assurance (QA) professional to determine if the proposed design meets the initial set of business goals.

- Implementation: The sixth phase is when the majority of the code for the program is written, and when the project is put into production by moving the data and components from the old system and placing them in the new system via a direct cutover [34].
- Maintenance [35]: The last phase is when end users can fine-tune the system, if they wish to boost performance, add new capabilities to meet additional user requirements.

3) *Theory of QR-Code*: The QR code system was invented in 1994 by Masahiro Hara from the Japanese company Denso Wave. The initial design was influenced by the black and white pieces on a Go board [36]. Its purpose was to track vehicles during manufacturing; it was designed to allow high-speed component scanning.

ISO/IEC 18004: 2015 defines the requirements for the symbology known as QR Code [37]. It specifies the QR Code symbology characteristics, data character encoding methods, symbol formats, dimensional characteristics, error correction rules, reference decoding algorithm, production quality requirements, and user-selectable application parameters.

The amount of data that can be stored in the QR code symbol depends on the data type (mode, or input character set), version (1, ..., 40, indicating the overall dimensions of the symbol, i.e., $4 \times \text{version number} + 17$ dots on each side), and error correction level. The maximum storage capacities occur for version 40 and error correction level L (low), denoted by 40-L.

TABLE I. CHARACTER REFERS TO INDIVIDUAL VALUES OF THE INPUT MODE/DATA TYPE

Input mode	Max. characters	Bits/char.	Possible characters, default encoding
Numeric only	7,089	31/3	0, 1, 2, 3, 4, 5, 6, 7, 8, 9
Alphanumeric	4,296	51/2	0-9, A-Z (upper-case only), space, \$, %, *, +, -, ., /, :
Binary/byte	2,953	8	ISO 8859-1
Kanji/kana	1,817	13	Shift JIS X 0208

From Table I, there are four input modes: 1) Numeric only: The maximum character is 7,089, and the minimum character is three and one third. 2) Alphanumeric: The maximum character is 4,296, and the minimum character is five and one half. 3) Binary/byte: The maximum character is 2,953 and the minimum character is 8. 4) Kanji/kana: The maximum character is 1,817 and the minimum character is 13.

4) *Waste management*: Solid waste refers to [38], the garbage generated from various activities in the community, such as homes, commercial establishments, business centers, shops, entertainment spot, fresh-food markets and institutions. The type of waste consists of organic waste: food scraps, leaf scraps, and grass clippings, etc.; recycled waste: glass, plastic,

aluminum, rubber, etc.; general waste: remnant of cloth, wood scraps; and other materials, including hazardous waste from the community. Solid waste management principles are as follows:

- Solid waste collection [39] is the collection of solid waste arising from various sources by considering the solid waste containers or trash cans, which can be classified and stored efficiently, as well as designated collection points for solid waste to facilitate the collection, save time and money.
- Transportation of solid waste [40] is the process of bringing the garbage that people from each household put in the solid waste bin to be transported for further disposal by solid waste truck. That should take into account the suitability of the area and the amount of solid waste that collects each day by demarcation of the solid waste collection route to be effective, save resources, and cost.
- Solid waste disposal [41], large-scale waste disposal technology, including composting system by decomposing organic matter through the biological process of microorganisms, the decomposition transforms them into minerals that are relatively stable. Sanitary landfill system, the solid waste is brought to the landfill in the area that has been prepared according to academic principles in terms of economic, social, environmental, engineering, architecture and public consent. There are measures to prevent potential impacts such as contamination of wastewater. In addition, measures must be taken to prevent flooding, smell pollution and impact on the landscape. The incinerator system is a solid waste management method in the incinerator that has been properly designed and constructed because it may cause air pollution, such as small dust particles, various toxic gases like sulfur dioxide, etc. Therefore, it is necessary to have an air pollution control system and to prevent the air passing through the flue into the atmosphere exceeds the air quality standard from the furnace.

III. METHODOLOGY

Develop an application for a waste management via the QR Code system. The principles of the System Development Life Cycle (SDLC) were used as follows.

1) *Requirement definition stage*: Studying documents and related research, as well as studying the problems and situations in the community regarding the classification of waste in order to determine the value of the purchase in the group, including waste management of Ban Yang sub-district communities to analyze and define the application development model.

2) *System analysis stage*: Analyzing and studying application patterns of applications in waste management via the QR code system to be consistent with community waste management activities; designed a context diagram containing information about the administrator, community

administrators, and community group members by defining the related work processes, including junk data management, member information management, QR code data management, reward information management, information management, and report display.

3) *System design stage*: Planning and defining work procedures to be consistent with the timeline, budget, and scope of the application for a waste management via the QR Code system. Determining the responsibilities for the working team and coordinated with people in the community involved, as is shown in Fig. 1.

4) *System development stage*: Developing the application for a waste management via the QR Code system, divided into two parts: 1) The user section, consists of the login screen, QR code scanning by managing the validation and designing with JAVA language and XML language, as well as managing database storage with MySQL engine. 2) The administrator section, manages application usage data to perform QR code scanning and the information management on the website, display Press release information, activity pictures, and display the household coordinates of the group members.

5) *System testing stage*: Testing the system using the application for a waste management via the QR Code system to check the operation of the system before it was put into actual use along with improving the syntax of the application's instruction set to meet the needs of users.

6) *System installation stage*: Executing the installation of the system to use the application for a waste management via QR Code system by using it in parallel with the existing community waste bank model and creating relevant documents such as a QR code diagram of the waste classification and application manual.

7) *System maintenance stage*: Improving the error from the operation of the application for a waste management via QR Code system including tracking the results of use and listening to suggestions on problems that arose in order to improve the system further.

After finishing the developing of an application for a waste management via the QR Code system, application technology dissemination is presented as is shown in Fig. 2.

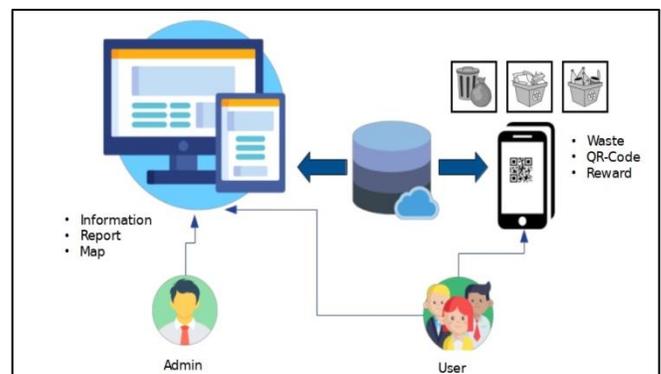


Fig. 1. Designing the Application for a Waste Management via the QR Code System.



Fig. 2. Presentation for Application Technology Dissemination.

IV. RESEARCH RESULT

A. Results of Analysis

Designing and developing of the application for a waste management via QR Code system use data from the study of problems in waste management in Ban Yang sub-district communities, Muang Buriram district, and Buriram province. These were analyzed and created a context diagram to show an overview of the system performance. The scope of the work studied in relation to the external environment of the system consists of the user menu. This can be checked by the member's points to be redeemed for rewards. Admin data management menu can manage member information, household coordinates, press release information, and QR code information, as is shown in Fig. 3.

B. Results of a Study

1) *Results of a study* on the quality of the application for a waste management via QR Code system by three experts and analyzed the data using mean and standard deviation according to the Likert method [42]. Details are shown in Table II.

From Table II, the results of the study on the quality of the application for a waste management via QR Code system by three experts revealed that the overall quality of the application was at a high level. The application design aspect was at the highest level. In terms of application usage and information security aspects were at a high level.

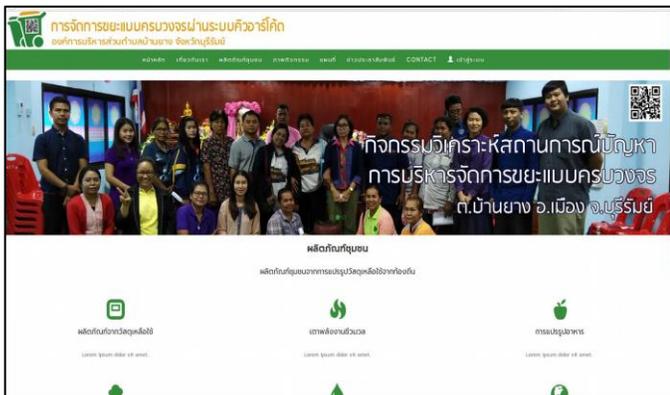


Fig. 3. Admin Data Management Menu.

TABLE II. THE RESULTS OF THE STUDY ON THE QUALITY OF THE APPLICATION FOR A WASTE MANAGEMENT VIA QR CODE SYSTEM

Appraisal item	\bar{x}	S.D.	Quality level
1. Application usage aspect	4.33	0.01	high
2. Application design aspect	4.46	0.30	the highest
3. Information security aspect	4.44	0.00	high
Total	4.41	0.10	high

Following Fig. 4 is a prototype of the application for a waste management via the QR-Code System.

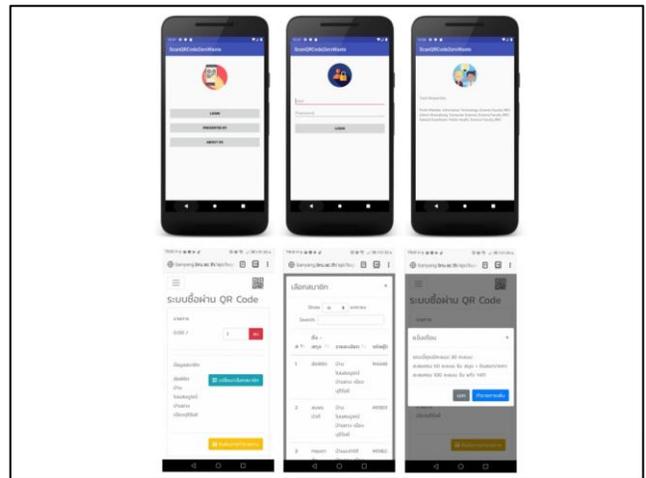


Fig. 4. Prototypes of the Application.

The system was installed and the developed application was tested with 388 people of sample groups, then the satisfaction study results were analyzed by basic statistical values compared to the criteria and summarized [43]. Details are shown in Table III.

TABLE III. THE RESULTS OF THE SATISFACTION ASSESSMENT OF THE APPLICATION USERS

Appraisal item	\bar{x}	S.D.	Satisfaction level
Application design aspect			
1. The font size was appropriate.	4.61	0.49	the highest
2. The topics used in the application were categorized appropriately.	4.27	0.62	high
3. Graphics used in the application were appropriate.	4.25	0.45	high
4. The use of color tones was appropriate and gorgeous.	4.23	0.42	high
5. The layout of the application components was uncomplicated to understand.	4.03	0.25	high
Performance in use aspect			
6. Users could scan the QR code easily and conveniently.	4.76	0.43	the highest
7. Users could manage information conveniently.	4.53	0.49	the highest
8. Users could speedily access information within the application.	4.91	0.45	the highest
9. The application performance overview was appropriate.	4.23	0.42	high
Total	4.42	0.45	high

The results of the satisfaction assessment of the users of the application for a waste management via QR Code system revealed that the overall was at a high level. When these were considered for each item, users were satisfied with the speed access to inform within the application at the highest level with the average 4.91 followed by users who could scan a QR code easily and conveniently with the average of 4.76. The font size was appropriate with the average of 4.61. The users could manage information conveniently with the average of 4.53. The topics used in the application were categorized appropriately with the average of 4.27. The graphics used in the application were appropriate and the application performance overview was appropriate with the average of 4.23. And the layout of the application components is uncomplicated to understand with the average of 4.03.

V. DISCUSSION AND CONCLUSION

From the research and development of the application for a waste management via the QR Code system, the related document was studied based on the System Development Life Cycle [44] through a process of participation in Ban Yang Sub-District Community, Mueang Buriram District, Buriram Province. The study analyzed the approach to waste management by applying application technology to store data and related information, divided into two parts: the application part was used in the QR scanning management to classify the types of waste in order to buy waste directly from the community group members by using the motivation to collect 100 points, and would receive a reward. That could check the history of buying waste of the group members, display a list of QR codes and report on the amount of waste by period [45]. This has been consistent with the research on QR CODE in Thailand and Application of QR Code Technology in the Hospitals in Thailand 4.0 by using QR code technology in service, thereby reducing errors streamlines the operations of various procedures, reducing redundant management, and making the service users get the most benefit.

The results of the quality assessment of the application by the experts were at a high level, which have been consistent with the research on Using QR-Code Technology for Public Relations of Community Enterprises Antique Hand-Woven Cloth Group, Padee Pha Tho, Tambon Bo Suak, Mueang district, Nan province. However, the users of the application for a waste management via QR Code system were satisfied with the system at a high level. The researchers in [46], in terms of application design and data accuracy [47] have been in line with the research on The Application of Two Dimension Barcode Technology for Providing Tourist Information Services at Tourism Destination Case Study: Doi Suthep Temple, Chiang Mai, which was designed and developed an application for information display that met the needs of users. That was easy access to information and convenient in service.

However, this should be further developed regarding the variety of applications in other uses, such as household location data in order to link the data to make it accessible to all. Local governments and community leaders should promote and support it as a learning area to be a role model in community waste management and to publicize knowledge and technology to nearby areas and people who are interested.

ACKNOWLEDGMENT

This research project was funded by the National Research Council of Thailand (NRCT). Thank you to the Research Progress Assessment Committee, Research and Development Institute, Buriram Rajabhat University. Thank you to the administrators of the Ban Yang Subdistrict Administrative Organization, Muang Buriram District, Buriram Province, officials, community leaders, seniors, youths, and people in the area for participating in research activities.

REFERENCES

- [1] Ban Yang Subdistrict Administrative Organization. (2018). Development Plan 4 years 2017 – 2021. Buriram Province: Policy and Plan Analysis, Ban Yang Subdistrict Administrative Organization, Mueang District, Buriram Province.
- [2] Boonlong, P. (2017). Village Headman. Buriram Province: Baanyang District, Mueang Buriram. Interview.
- [3] Suwanmala, C. (2010). Decentralization and Thailand Reform. Bangkok: Center for Innovation and Local Good Governance (CLTG), Faculty of Political Science, Chulalongkorn University.
- [4] Mark A Russo CISSP-ISSAP ITILv3. (2019). The Agile/Security Development Life Cycle (A/SDL): Integrating Security Functionality into the SDLC (2nd ed). Cybersentinel, LLC.
- [5] Guimaraes, T. (1985). A study of application program development techniques. *Communications of the ACM*, 28(5), 494–499. <https://doi.org/10.1145/3532.3534>.
- [6] Malcolm, D. G., Roseboom, J. H., Clark, C. E., & Fazar, W. (1959). Application of a Technique for Research and Development Program Evaluation. *Operations Research*, 7(5), 34–47. <https://doi.org/10.1287/opre.7.5.646>
- [7] Calder, B., Phillips, L., & Tybout, A. (1981). Designing Research for Application. *Journal of Consumer Research*, 8(2), 197–207. <https://doi.org/10.1086/208856>
- [8] Perkins, D., & Zimmerman, M. (1995). Empowerment theory, research, and application. *American Journal of Community Psychology*, 23(5), 569–579.
- [9] Casteleyn, S., Daniel, F., Dolog, P., & Matera, M. (2009). *Engineering Web Application*. Springer-Verlag Berlin Heidelberg.
- [10] Bates, C. (2000). *Web Programming: Building Internet Applications*. Wiley.
- [11] Jobe, W. (2013). Native Apps Vs. Mobile Web Apps. *International Journal of Interactive Mobile Technologies (IJIM)*, 7(4), 27–32. <https://doi.org/10.3991/ijim.v7i4.3226>
- [12] Meirelles, P., Rocha, C., Assis, F., Siqueira, R., & Goldman, A. (2019). A Students' Perspective of Native and Cross-Platform Approaches for Mobile Application Development. *Computational Science and Its Applications – ICCSA 2019*. https://doi.org/10.1007/978-3-030-24308-1_47
- [13] Bernardes, T.F., Miyake, M.Y. (2016). Cross-platform mobile development approaches: a systematic review. *IEEE Lat. Am. Trans.* 14(4), 1892–1898.
- [14] Jhala, D. (2021). A Study on Progressive Web Apps as A Unifier for Native Apps and the Web. *International Journal of Engineering Research & Technology (IJERT)*, 10(5), ISSN (Online): 2278-0181.
- [15] Svensson, O., & Kåld, M. P. (2021). React Native and native application development. Faculty of Computer Science, Jönköping University.
- [16] Verma, N., Kansal, S., & Malvi, H. (2018). Development of Native Mobile Application Using Android Studio for Cabs and Some Glimpse of Cross Platform Apps. *International Journal of Applied Engineering Research*, 13(16), 12527–12530.
- [17] Huynh, M. Q., & Ghimire, P., & Truong, D. (2017). Browser App Approach: Can It Be an Answer to the Challenges in Cross Platform App Development?. *Journal of Information Technology Education: Innovations in Practice*, 16, 47–68.

- [18] Huynh, M. Q., Ghimire, P., Truong, D. (2017). Hybrid App Approach: Could It Mark the End of Native App Domination?. *Journal of Informing Science Institute*, 14, 49-65. <https://doi.org/10.28945/3723>.
- [19] Panhale, M. (2016). *Beginning Hybrid Mobile Application Development*. Apress, Berkeley, CA. Mark A Russo CISSP-ISSAP ITILv3, (2019), "The Agile/Security Development Life Cycle (A/SDLC): Integrating Security Functionality into the SDLC", Second Edition, Cybersentinel, LLC.
- [20] Kho'i, F. M., & Jahid, J. (n.d.). Comparing Native and Hybrid Applications with focus on Features. Faculty of Computing, Blekinge Institute of Technology.
- [21] Khandeparkar, A., Gupta, R., & Sindhya, B. (2015). An Introduction to Hybrid Platform Mobile Application Development. *International Journal of Computer Applications*, 118(15), 31-33.
- [22] Sahoo, R., & Sahoo, G. (2016). *Multimedia & Web Technology*. New Saraswati House (India) Pvt. Ltd.
- [23] Al-Fedaghi, S. (2011). Developing Web Applications. *International Journal of Software Engineering and Its Applications*, 5(2), 57- 68.
- [24] Rossi, G., Pastor, O., Schwabe, D., & Olsina, L. (2008). *Web Engineering: Modelling and Implementing Web Applications*. Springer-Verlag London Limited 2008.
- [25] Fraternali, P. (1999). Tools and approaches for developing data-intensive Web applications: a survey. *ACM Computing Surveys*, 31(3),227-263. doi.org/10.1145/331499.331502.
- [26] Paulson, L. D. (2005). Building rich web applications with Ajax. *IEEE*, 38(10), 14-17. <https://doi.org/10.1109/MC.2005.330>.
- [27] Wassermann, G., Yu, D., Chander, A., Dhurjati, D., Inamura, H., & Su, Z. (2008). Dynamic test input generation for web applications, *ISSTA '08*. Proceedings of the 2008 international symposium on Software testing and analysis, 249–260. <http://doi.org/10.1145/1390630.1390661>.
- [28] Everett, G.D., & McLeod, R. (2007). *Software Testing*. John Wiley & Sons, Inc.
- [29] Bryant, P., Howard, D., Lock, G., & Philander, Z. (2008). *Systems Analysis and Design*. Pearson Education South Africa.
- [30] Dixit J. B. (2007). *Structured System Analysis and Design*. Laxmi Publications.
- [31] Dennis, A., Wixom, B., & Roth, R.M. (2018). *Systems Analysis and Design (7th ed)*. Wiley.
- [32] Awad, E.M. (1992). *Systems Analysis and Design (2nd ed.)*. Galgotia Publications.
- [33] Valacich, J., & George, J. F. (2017). *Modern Systems Analysis and Design*, Global Edition (8th ed.). Pearson (US).
- [34] Singh, P.P., Kaur, S., & Sharma, S. (2006). *System Analysis and Design: Complete Course Book*. Deep & Deep Publications Pvt. Ltd.
- [35] Whitten, J.L., Bentley, L.D., & Barlow, V. (2005). *Systems Analysis and Design Methods*. McGraw-Hill Education.
- [36] Wave, D. (2018). QR Code development story, <https://www.denso-wave.com/en/technology/vol1.html>.
- [37] ISO/IEC 18004: 2015. (2018). Information technology — Automatic identification and data capture techniques — QR Code bar code symbology specification. <https://www.iso.org/standard/62021.html>.
- [38] Ravindraa, K., & Morb, S. (2019). E-waste generation and management practices in Chandigarh, India and economic evaluation for sustainable recycling. *Journal of Cleaner Production*, 221, 286-294.
- [39] Maletz, C.R., & Ziyang, D.L. (2018). *Source Separation and Recycling Implementation and Benefits for a Circular Economy*. Springer International Publishing AG.
- [40] Chandra, R. (2015). *Environmental Waste Management*. CRC Press.
- [41] Tchobanoglous, G., & Kreith, F. (2002). *Handbook of Solid Waste Management (2nd ed.)*. The McGraw-Hill Companies, Inc.
- [42] Likert, R. (2013). The Method of Constructing and Attitude Scale. In Reading in Fishbein, M (Ed.), *Attitude Theory and Measurement*, New York: Wiley & Son.
- [43] Ajzen, I. (1993). *New directions in attitude measurement*. Berlin: Walter de Gruyter.
- [44] Murch, R. (2012). *The Software Development Lifecycle – A Complete Guide*. Amazon Asia Pacific Holdings Private Limited.
- [45] Kittesh, A., & Kaeboodee, K. (2019). QR CODE in Thailand and Application of QR Code Technology in the Hospitals in Thailand 4.0. *Mahidol R2R e-journal*, 5(2), 51-59.
- [46] Hongsibsong, P., Turayot, P., & Kanlanon, P. (2021). Web Application Development for Waste Bank Management by MahaPho Community Participation Mueang district, Nan province. *The National Conference on Technology and Innovation Management*, 8(1), 73-86.
- [47] Kanchanawong, P., & Kanchanawong, A. (2018). The Application of Two Dimension Barcode Technology for Providing Tourist Information Services at Tourism Destination Case Study: Doi Suthep Temple, Chiang Mai. *Journal of Humanities and Social Sciences*, 9(17), 120-134.

Empirical Study of a Spatial Analysis for Prone Road Traffic Accident Classification based on MCDM Method

Anik Vega Vitianingsih¹

Informatics Departments
Universitas Dr. Soetomo

Faculty of Information and Communication Technology
Universiti Teknikal Malaysia Melaka
Surabaya, Indonesia. Melaka, Malaysia

Zahriah Othman², Safiza Suhana Kamal Baharin³

Faculty of Information and Communication Technology
Universiti Teknikal Malaysia Melaka, Melaka, Malaysia

Aji Suraji⁴

Department of Civil Engineering
University of Widyagama Malang, Malang, Indonesia

Abstract—Spatial analysis techniques are widely used as an effective approach for prone road traffic accident classification. This paper will present the results of empirical behavioral testing on the spatial analysis for prone road traffic accident classification using the Multicriteria Decision Making (MCDM) method. The performance of MCDM is compared on arterial and collector road types processed with multicriteria parameters. MCDM was chosen because it can be used as a decision making based on an alternative selection with many criteria. Empirical tests of the MCDM method used include Weighted Sum Model (WSM), Weighted Product (WP), Simple Additive Weighting (SAW), Weighted Product Model (WPM), Multi-Attribute Utility Theory (MAUT), Technique for Others Reference by Similarity to Ideal Solution (TOPSIS), and Analytical Hierarchy Process (AHP). The multicriteria parameter weight values are based on expert judgment and the Fuzzy-AHP method (EJ-AHP), which comprises volume-to-capacity ratio (VCR), international roughness index (IRI), vehicle type, horizontal alignment, vertical alignment, design speed, and shoulder. Then, the performance of the models was compared to determine the value of accuracy, precision, recall, and F1-score as decision-making on the prone road traffic accident classification using Multicriteria Evaluation Techniques (MCE). The empirical test results on arterial roads show that the SAW and TOPSIS methods have the same performance and are superior to other methods, with an accuracy value of 63%. However, the results on the collector road type show that the accuracy value of the AHP method outperforms other methods with an accuracy value of 70%.

Keywords—Spatial Analysis; GIS; prone road traffic accident; MCDM Model; WSM; WP; SAW; WPM; MAUT; TOPSIS; AHP

I. INTRODUCTION

The rate of road traffic accidents (RTA) that results in deaths increases every year. Data for 2004-20130 states that RTA is the leading cause of death, which is ranked 9th in the world; WHO estimates that in 2030 the RTA will increase to the 5th rank if there are no efforts to overcome this problem [1]. The number of deaths due to RTA annually reached 1.35 million worldwide in 2016 [2].

The case study for the spatial analysis for prone road traffic accident classification in the discussion of this paper is on the type of arterial and collector roads in the Province of East Java, Indonesia, which is one of the areas with a very high accident-prone. The Global Status Report on Road Safety 2018 states that in Indonesia, with a population number is 261,115,456 people in 2016, the number of deaths due to RTA reached 31,282 million people [2]. The accident factors include 69.70% due to the human factor, 21.21% due to road facilities, and 9.09% due to road infrastructure factor (Komite Nasional Keselamatan Transportasi, 2016). In 2010 the United Nations General Assembly declared a Decade of Action for Road Safety year 2011-2020 aimed at stabilizing the level of fatality of global casualties by increasing activities undertaken at national, regional, and global scales [1][3][4]. The spirit of the Road Safety Action Declaration 2011-2020 is in line with the mandate of Law Number 38 of 2004 [5], Number 34 of 2006 [6] concerning roads, and Law Number 22 the Year 2009 concerning road traffic and transportation [7] to prepare a National General Plan for Road Safety 2011-2035 in Indonesia [8] as outlined in and Regulation of the President of the Republic of Indonesia Number 2 of 2012 concerning the national transportation safety committee [9].

Spatial data modeling (GIS-spatial analysis) is a part of multicriteria decision-making (GIS-MCDM). Spatial analysis in geographic information systems (GIS) is the process of developing artificial intelligence (AI) formulations by combining geo-referenced data (spatial data) with multicriteria parameters as value assessment attribute data (decision-makers preferences and uncertainty) to obtain the appropriate information in georeferencing-based decision making. GIS is commonly regarded as a technology capable of integrating, storing, manipulating, analyzing, and displaying spatial data and attribute data for decision-making and decision-supporting operations [10]. The MCDM method provides a collection of procedures and AI algorithms for formulating decision-making problems, designing, evaluating, and prioritizing alternative decisions [11][12]. This empirical study aims to analyze the sensitivity of the methods tested through spatial data modeling with MCE. MCE evaluates the methods [13] by

testing the extent to which the values of accuracy, precision, recall, and F1 scores when multicriteria parameters systematically vary on various interests.

The characteristic of GIS-Spatial MCDM is to determine the weighting of the spatial datasets used for spatial data modeling. The literature study [14]–[18] summarizes several issues related to GIS-spatial relationship modeling for prone-roads classification traffic accidents (PRTA) using the MCDM method. First, GIS-Spatial relationship modeling using multicriteria decision-making methods (GIS-Spatial MCDM) is a spatial analysis process to perform spatial data modeling that involves multicriteria parameters from the expert judgment in spatial decision making. The spatial analysis involves multicriteria parameters in building software GIS for spatial decision-making based on combined theory, methods, and measurement tools from expert judgment [19][20]. An expert judgment is required to validate the spatial dataset parameters used [19].

Many researchers give parameter-weighted values only from the point of view of expert judgment. The expert judgments give a subjective and objective risk and bias in the evaluation process for weighting and parameter priority scale [21]. The weighted value given to the multicriteria parameter will impact the accuracy of the spatial data modeling results for PRTA classification [22][23]. Many researchers claim that the multicriteria parameters used are effective and capable of determining the PRTA classification [24]–[32].

Secondly, classification techniques are needed to produce accurate spatial data modeling without overlapping interests and to avoid overfitting problems, and the deep-neuro-fuzzy classification method is used for road weight measurement [33]. The analytical hierarchy process (AHP) method can improve the road safety audit technique used to identify and prioritize black spots in the absence of statistical data of accidents that were not recorded correctly. [34]. The AHP method is used to perfect the weighting value generated from literature studies and expert assessments [31][35] by determining the priority scale ranking of the parameters using the random forest (RF) [31] method, preference ranking organization method for enrichment of evaluations (PROMETHEE), and VlseKriterijuska Optimizacija I Komoromisno Resenje (VIKOR) method [35].

The literature study shows that the MCDM method such as SAW and AHP methods and Fuzzy AHP for determining the weight can help decisions process in Road Safety Analysis (RSA) such as road management prioritization and provide mitigating actions against the most vulnerable to accidents. In another literature study, the TOPSIS classification model is used to manage road safety to reduce the number of traffic accidents by knowing the position of a road safety study based on various quantitative and qualitative criteria [36]. Besides that, the simple ranking (SR) method and the empirical Bayes (EB) combine the type and severity of the accident to the data series in Australia, then proposed to evaluate alternative indicators with multiple criteria parameters for the identification of accident-prone road (blackspot). The SR and EB method is used to calculate the value of accidents by type

of case, societal cost of any accident, and the crash prediction models using data series [37].

GIS is region specific [38][39], where 96% [14] of research uses private spatial datasets with small dataset characteristics. The challenge for small datasets is at the data pre-processing stage to produce optimal performance on the AI method used. The GIS-Spatial MCDM approach is proposed in this study based on the characteristics of the MCDM model based on multi-criteria parameter weighting. The weighting that has been carried out by expert judgment will be combined with the AHP computational method (EJ-AHP) as a means of measuring the resulting weight value.

Based on the review of these literature studies, however, no research studies specifically for evaluating and comparing through an empirical study approach in the spatial analysis using the MCDM method (WSM, WP, SAW, WPM, MAUT, TOPSIS, and AHP methods) for prone road traffic accidents. The classification of multicriteria parameter weight values is based on expert judgment and the AHP method (EJ-AHP). Therefore, this study proposed a combination of expert judgment and Fuzzy-AHP (EJ-Fuzzy-AHP) to produce weighting values in spatial datasets and provide the appropriate parameter priority scale values. Fuzzy-AHP has a procedure following decisions that involve expert judgment, so it can be used to combine data knowledge in Fuzzy-AHP with expert judgment. Then, these weighting values were used in the MCDM method for GIS-based spatial modeling for PRTA classification based on multicriteria parameters, namely speed design, volume/capacity ratio (V/C Ratio) [40], the width of the road, the number of lanes, road shoulders, median strip, horizontal alignment, vertical alignment, road condition [40], and vehicle type.

The discussion structure in this paper includes section II: which discusses multicriteria parameters through the description of spatial datasets, section III: which discusses research methodology; section IV: which describes results and discussion; and section V: which discusses the conclusion and future works directions.

II. SPATIAL DATASETS

The spatial dataset parameters in this study were obtained from private data sources, so a decision-making model using GIS-Spatial MCDM is proposed. The GIS decision-making system is used for specific regions case studies in which 96% of researchers use private data types on specific regional case studies [14] with multicriteria parameters from the expert judgments. The MCDM method is applied to making decisions through management priority ranking related to existing or specific region-specific planning policies [41]. The MCDM method is one of the right approaches to deal with the problem of the PRTA classification because it uses several road and environmental criteria, both quantitative and qualitative; MCDM is related to the results of decision making for planning that involves stakeholders [42].

According to the Republic of Indonesia Law No.38 of 2004 article 8, the type of road based on its function is divided into 4 (four), namely arterial roads, collector roads, local roads, and environmental roads. The arterial road is a public

road with the number of access roads is limited efficiently, works to serve the main transport which connects provincial capitals with the characteristics of long-distance travel dan high average speeds [43]–[45]. The collector road is a public road that works to serve vehicle which connects between regency capitals with medium distance travel characteristics, medium average speed, and a limited number of entrances [43][44]. Local roads are public roads with an unlimited number of access roads that serve local transport, which link the sub-district cities with short-distance travel and low average speeds [43][44]. The environmental road is a public road that serves environmental transportation between villages with short-distance travel characteristics and low average speed [43][44].

In this study, the spatial analysis datasets include the types of primary arterial networks and primary collector networks since both types of roads are the main roads that supply sufficient datasets for this study. The case study involves the research objects in Indonesia with data retrieval from the National Road Development Center, Police Corps and Traffic Police of Indonesia, Traffic Corps National Police, and Transportation Department. Descriptions of the spatial datasets on the multicriteria parameters used are shown in Table I were to the range and score for Spatial Datasets from:

- The Directorate General of Highways Standard Specifications for Geometric Design of Urban Roads. Ministry of Public Works, Directorate General of Highways, Jakarta 1992.
- Indonesian Highway Capacity Manual (IHCM), 1997
- TRB Highway Capacity Manual. Transportation Research Board Special Report 209; Washington D.C. USA 1985. Revised 1994.
- SNRA Manual on Calculation of Capacity, Queues, and Delay in Traffic Facilities (in Swedish). Swedish National Road Administration Report TV 131, 1977.

Geospatial Datasets comprised spatial data needs for a base map (layer) and attribute data requirements for multicriteria parameters that were utilized for spatial analysis of PRTA classification. The data requirements used in this study used private data types from the National Road Implementation Center, East Java Bali, Indonesia. Spatial datasets include:

1) *The base map*: consists of attributes road number, suffix, road names, length of roads (km), and road function.

- a) *Arterial primary networks*
- b) *Collector primary networks*

2) *Multicriteria parameters*

a) *Volume-to-capacity ratio (VCR)*: to measure the overall service quality provided. If the VCD is high, it indicates a high risk of accidents.

b) *International Roughness Index (IRI)*. Condition of the pavement. If the IRI is heavily damaged, the likelihood of an accident increases significantly.

c) *Vehicle Type*: vehicle types 2/1 UD, 2/2 UD, 4/2 UD, 4/2 D, and 6/2 can pass through arterial or principal collector roads.

d) *Horizontal alignment (HA)*. Projection of the axis of the road for roads without a median or the projection of the inner edge of the pavement for roads with a median. If the horizontal alignment is sharp, the potential for accidents is high.

e) *Vertical alignment (VA)*: the intersection of the vertical plane with the road pavement surface through the road axis for 2-speed 2-way roads or through the inner edge of each pavement for roads with a median. If the vertical alignment is high, the potential for accidents is high.

f) *Design speed (Vr)*: The vehicle speed can be achieved safely when running without interruption. If the speed is high, then the accident potential is high.

g) *Shoulder*: The lane is located side by side with the traffic lane. If the shoulder there isn't, then the potential for an accident is high.

TABLE I. SPATIAL DATASETS PARAMETERS

Arterial Road			
Parameters	Range	Description	EJ Scoring
VCR (%)	$VCR \geq 0.85$ && $VCR < 1.00$	The condition reaches capacity with 2000 units of passenger cars (pcu/hour), 2 directions.	5
	$VCR \geq 0.70$ && $VCR < 0.85$	The conditions approach unsteady flow with traffic volume reaching 85% of the capacity, namely 1700 units of passenger cars (pcu/hour), 2 directions.	4
	$VCR \geq 0.45$ && $VCR < 0.70$	Traffic flow conditions are still stable, with traffic volume reaching 70% of capacity (pcu/hour), 2 directions	3
	$VCR \geq 0.20$ && $VCR < 0.45$	The start of a stable flow condition with traffic volume reaching 45% of the capacity is 900 units of passenger cars (pcu/hour), 2 directions.	2
	$VCR < 0.20$	The conditions free flow with traffic volume reaching 20% of the capacity, namely 400 units of passenger cars (pcu/hour), 2 directions.	1
IRI (m/km)	$IRI \geq 12$	Heavy Damage	4
	$IRI \geq 8$ && $IRI < 12$	Light Damage	3
	$IRI \geq 4$ && $IRI < 8$	Moderately	2
	$IRI < 4$	Good	1
HA(rad/km)	$HA \geq 3.50$	Poor	3
	$HA \geq 0.25$ && $HA < 3.50$	Fair	2
	$HA < 0.25$	Good	1
VA	$VA \geq 45$	Poor	3

Arterial Road			
Parameters	Range	Description	EJ Scoring
(m/km)	$VA \geq 5$ && $VA < 45$	Fair	2
	$VA < 5$	Good	1
V_r (km/jam)	$V_r \geq 100$	Traffic speed more than 100 kilometres per hour.	6
	$V_r \geq 80$ && $V_r < 100$	Traffic speed is more than 80 kilometers per hour.	5
	$V_r \geq 65$ && $V_r < 80$	Traffic speed more than 65 kilometres per hour.	4
	$V_r \geq 60$ && $V_r < 65$	The speed limit is reduced to 60 kilometers per hour.	3
	$V_r \geq 50$ && $V_r < 60$	The average pace of traffic is approximately 50 kilometers per hour.	2
	$V_r < 50$	Traffic moving at a speed of fewer than 50 kilometers per hour.	1
Road Type	2/2 UD	The traffic road is a two-lane two-way without a median (2/2 UD)	5
	4/2 UD	The traffic road is a four-lane two-way without a median (4/2 UD)	4
	4/2 D	The traffic road is four lanes two-way with a median (4/2 D).	3
	6/2 D	The traffic road has six two-way lanes with a median (6/2 D).	2
	2/1 UD	The traffic road has two lanes with no median (2/1 UD).	1
Shoulder	No	No, there is no roadside shoulder.	2
	Yes	Yes, the roadside shoulder is available.	1
Collector Road			
Parameters	Range	Description	EJ Scoring
VCR (%)	$VCR \geq 0.90$ && $VCR < 1.00$	The condition reaches capacity with 2000 units of passenger cars (pcu/hour), 2 directions.	5
	$VCR \geq 0.75$ && $VCR < 0.90$	The conditions approach unstable flow with traffic volume reaching 90% of the capacity, namely 1800 units of passenger cars (pcu/hour), 2 directions.	4
	$VCR \geq 0.50$ && $VCR < 0.75$	Traffic flow conditions are still stable, with traffic volume reaching 75% of capacity (pcu/hour), 2 directions	3
	$VCR \geq 0.30$ && $VCR < 0.50$	The start of a stable flow condition with traffic volume reaching 50% of the capacity is 1000 units of passenger cars (pcu/hour), 2 directions.	2
	$VCR < 0.30$	The conditions free flow with traffic volume reaching 30% of the capacity, namely 600 units of passenger cars (pcu/hour), 2 directions.	1
IRI (m/km)	$IRI \geq 12$	Heavy Damage	4
	$IRI \geq 8$ && $IRI < 12$	Light Damage	3
	$IRI \geq 4$ && $IRI < 8$	Moderately	2

Arterial Road			
Parameters	Range	Description	EJ Scoring
HA (rad/km)	$IRI < 4$	Good	1
	$HA \geq 3.50$	Poor	3
	$HA \geq 0.25$ && $HA < 3.50$	Fair	2
VA (m/km)	$HA < 0.25$	Good	1
	$VA \geq 45$	Poor	3
	$VA \geq 5$ && $VA < 45$	Fair	2
V_r (km/jam)	$VA < 5$	Good	1
	$V_r \geq 100$	Traffic speed more than 100 kilometres per hour.	6
	$V_r \geq 90$ && $V_r < 100$	Traffic speed is more than 80 kilometers per hour.	5
	$V_r \geq 75$ && $V_r < 90$	Traffic speed is more than 65 kilometers per hour.	4
	$V_r \geq 60$ && $V_r < 75$	The speed limit is reduced to 60 kilometers per hour.	3
	$V_r \geq 50$ && $V_r < 60$	The average pace of traffic is approximately 50 kilometers per hour.	2
Road Type	$V_r < 50$	Traffic moving at a speed of fewer than 50 kilometers per hour.	1
	2/2 UD	The traffic road is a two-lane two-way without a median (2/2 UD)	5
	4/2 UD	The traffic road is a four-lane two-way without a median (4/2 UD)	4
	4/2 D	The traffic road is four lanes two-way with a median (4/2 D).	3
	6/2 D	The traffic road has six two-way lanes with a median (6/2 D).	2
Shoulder	2/1 UD	The traffic road has two lanes with no median (2/1 UD).	1
	No	No, there is no roadside shoulder.	2
Shoulder	Yes	Yes, the roadside shoulder is available.	1

III. RESEARCH METHODOLOGY

The proposed MCDM experiment procedure in Fig. 1 has major differences from the existing framework [46]–[48]. Fig. 1 describes the proposed MCDM experiment procedure, namely:

- The requirement gathering a primary data set as a base map to determine the category of roads to be studied. This research uses private data types. The base maps used include primary arterial and primary collector networks.
- Attribute data for the multicriteria parameters used is based on an assessment by expert judgment, including VCR, IRI, vehicle type, horizontal alignment, vertical alignment, design speed, and shoulder. The data requirements are described in Table II.
- Conduct a literature study related to the multicriteria parameters used in each road category based on expert judgment assessment with the results in Table II.

- Mathematical modeling for spatial data analysis through empirical study for PRTA classification based on the MCDM method using WSM, WP, SAW, WPM, MAUT, TOPSIS, and AHP methods. In this case, the data pre-processing process will be carried out for the classification analysis process, that is:
 - Determine the priority weight of the parameters using the AHP method.
 - Determine the multiclass classification range obtained from the final value of the results of mathematical modeling on the MCDM method using the Guttman Scale. This process is carried out because there is no standardized assessment from expert judgment regarding the value of the PRTA classification range based on the multicriteria parameters used.
- The results of the multiclass classification will be validated through the value of accuracy and F1 score, which the F1 score is derived from the precision and recall.

A. The Priority Weight of the Parameters

Spatial decision-making based on multicriteria parameters is almost always faced with the problem of determining the

level of importance or influence between parameters. Decision-makers will weigh each parameter based on the importance or influence between these variables, which is usually done by expert judgment. The AHP method can solve the complex multicriteria parameter problems into a hierarchical unit. The hierarchy represents a complex problem in a multilevel structure, where the first level is the goal, followed by the factors level, criteria, sub-criteria, and the last level of alternatives. With a hierarchy, complex problems can be described in groups which are then arranged into a hierarchical form so that problems will appear more structured and systematic.

How to overcome the biases from the weighting given by expert judgment overdue of various factors of interest, then the decision-maker can perform parameter weighting using the AI method. The AHP is a pairwise comparison method through an analytic hierarchy process. The parameter weights are determined by normalization through the eigenvectors associated with the maximum eigenvalues in the unit ratio matrix. The weighting between parameters in this study is accomplished by the AHP method approach based on flow depicted in Fig. 2.

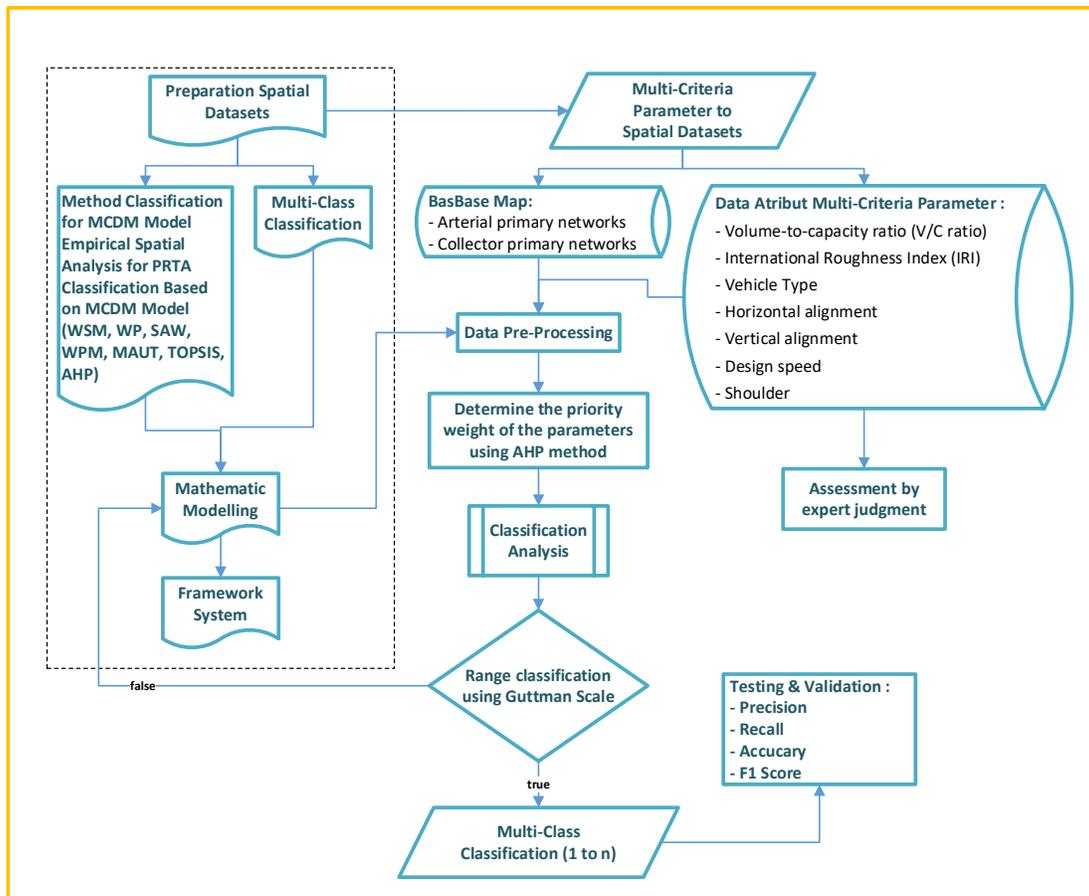


Fig. 1. Proposed MCDM Experiment Procedure.

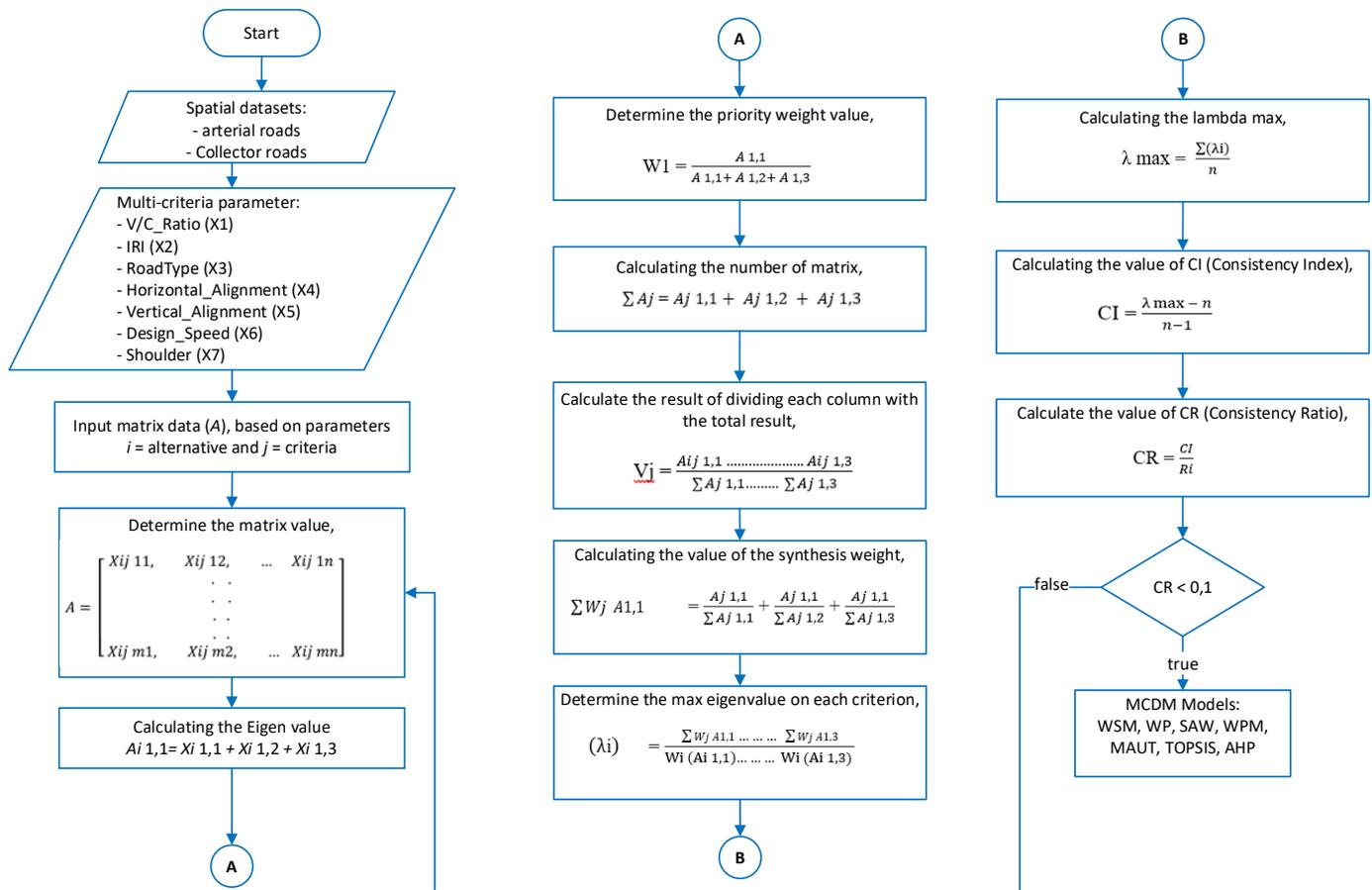


Fig. 2. The Priority Weight of the Parameters using the AHP Method.

Which consists of the following:

Step 1: Input the spatial datasets based on the base map used, namely the arterial and collector roads network.

Step 2: Input data multicriteria parameters, provide the labeling of the parameters used, namely VCR (X1), IRI (X2), Road Type (X3), HA (X4), VA (X5), Vr (X6), and Shoulders (X7).

Step 3: Input matrix data (A), based on parameters. Where i variable is an alternative, and j variable is criteria.

Step 4: Determine the multicriteria parameters matrix value (A) using pairwise comparison based on Eq. (1). A value is assigned to each criterion in accordance with the specifications of the hierarchical structure based on the number of multicriteria parameters present.

$$A = \begin{bmatrix} X_{ij 11}, X_{ij 12}, \dots X_{ij 1n} \\ \vdots \\ \vdots \\ \vdots \\ X_{ij m1}, X_{ij m2}, \dots X_{ij mn} \end{bmatrix} \quad (1)$$

The recommended values for creating a pairwise comparison matrix, where the values referring to Table III [49].

TABLE II. THE RECOMMENDED VALUES FOR CREATING A PAIRWISE COMPARISON MATRIX

Value	Description
1	Equally important (equal)
3	A little more important (slightly)
5	More importantly, with a strong type (strongly)
7	More importantly, with a very strong type (very strong)
9	More important to the extreme (extreme)

Step 5: Calculate the eigenvalues of each element in each pairwise comparison matrix. The eigenvalues are the weight of each element used to determine the priority of items in each hierarchical structure. Operate for adding values to each column in question to obtain the normalization of the matrix, based on Eq. (2).

$$A_{i 1,1} = X_{i 1,1} + X_{i 1,2} + X_{i 1,3} \dots X_{i 1,n} \quad (2)$$

Step 6: Calculate the priority weight value of the parameter (Wi) using Eq. (3), adding up each column's values in the pairwise comparison matrix, then dividing each value in the column by the total number of related columns obtain a normalized matrix. Where, $\sum A_{ij}$ is the number of matrices,

$$W_i = \frac{A_{i 1,1} \dots A_{i 1,5}}{\sum A_{ij}} \quad (3)$$

Step 7: Calculate the result of divide each column by the result of the total number (V_j) using Eq. (4), where $\sum A_j = A_{j,1,1} \dots \dots A_{j,1,5}$. Addition the values of each row and divide them with the number of elements to get the average value.

$$V_j = \frac{A_{ij,1,1} \dots \dots \dots A_{ij,1,5}}{\sum A_{j,1,1} \dots \dots \dots \sum A_{j,1,5}} \quad (4)$$

Calculate the value of the synthesis weight (W_j) to j using Eq. (5). Where $\sum W_j$ is the result of adding V_j to calculate the weight value of the synthesis.

$$\sum W_j = V_{j,1,1} \dots \dots \dots V_{j,1,5} \quad (5)$$

Step 8: Determine the max eigenvalue (λ_i) on each criterion using Eq. (6). Where, $\sum(\lambda_i)$ is $(\lambda_i) A_{1,1} \dots \dots (\lambda_i) A_{1,5}$.

$$(\lambda_i) = \frac{\sum W_j A_{1,1} \dots \dots \sum W_j A_{1,5}}{W_i (A_{i,1,1}) \dots \dots W_i (A_{i,1,5})} \quad (6)$$

Calculate the Lambda max (λ max) using Eq. (7). Where, $\sum(\lambda_i)$ is the total value of the sum of the eigenmax max, and n variable is the number of criteria.

$$\lambda \text{ max} = \frac{\sum(\lambda_i)}{n} \quad (7)$$

Step 9: Check the consistency of the hierarchy by calculating the value of the consistency ratio using the consistency index (CI) using Eq. (8), where:

- If the CI value 10%, then the consistency ratio is correct
- If the CI value is > 10%, then the consistency ratio is wrong, so data assessment must be corrected and reviewed.

$$CI = \frac{\lambda \text{ max} - n}{n - 1} \quad (8)$$

The consistency ratio (CR) value can be calculated using Eq. (9). Where R_i is a random index value determined by the hierarchy structure, as described in Table III, where:

- If $CR < 0.1$, then the level of consistency shown is quite rational in the pair comparison matrix.
- If $CR > 0.1$, it indicates an inconsistent assessment of the pair comparison matrix.

$$CR = \frac{CI}{R_i} \quad (9)$$

B. Weighted Sum Model (WSM)

WSM is a simple method that is extensively used in decision-making on single-dimensional problems. Attribute normalization is done by altering the value of the numeric column in the data set to the same scale to obtain a balance on the overall attribute value. Spatial data modeling with the WSM method is an approach to determining the weight of the priority value of each parameter of the attribute parameter, then multiplying with the data of each attribute to take a high alternative value as a solution [50]–[53].

The following sequence of the steps spatial analysis process for the PRTA classification using the WSM method:

Step 1: Follow the steps in the flow in Fig. 2 to define the criteria used as a benchmark for solving the problem and determine the priority of parameter weights.

Step 2: Calculate the priority values for each layer dataset by using the matrix of Eq. (10) [51].

$$A_i^{WSM-score} = \sum_{j=1}^n W_j X_{ij}; \text{ for } i = 1,2,3, \dots m \quad (10)$$

Where, $A_i^{WSM-score}$ is potential WSM score, X_{ij} a variable is an alternative to the i data score based on the j relative weight criterion, and W_j a variable is the j relative weight criterion

Step 3: Determine the range of PRTA classification values using the Guttman scale based on Eq. (11).

$$I = R / K \quad (11)$$

Where the I variable is the interval range, the R variable is the result of the calculation of the highest scores value of A_i minus the lowest score value of A_i , and the K variable is the number of alternatives. The alternative assessment criteria for the PRTA classification are obtained from the result of the calculation of the highest scores value of A_i minus the value of I variable as shown in the result of Eq. (12).

$$\begin{cases} \text{PRTA, if } A_i \geq \text{the scale range} \\ \text{Non_PRTA, if } A_i < \text{the scale range} \end{cases} \quad (12)$$

C. Weighted Product (WP)

The WP method is a decision support system that connects attribute ratings through multiplication operations to be raised to the power of the appropriate attribute weights. The normalization process to handle different units of measurement is done through multiplication operations on attribute ratings [54].

The following sequence of the steps spatial analysis process for the PRTA classification using WP method:

Step 1: Follow the steps in the flow in Fig. 2 to define the criteria used as a benchmark for solving the problem and determine the priority of parameter weights. Determine the initial and final input to change the name of the input into a rating value and determine the weight of each criterion. Improve the weights of each criterion by adding up the weights of each criterion, followed by dividing the result of the sum of the weights of the criteria by the starting weight of each criterion divided by the result of the sum of the weights of the criteria.

Step 2: Calculate the normalization value using Eq. (13) to get the alternative preference value of each criterion represented by the vector S_i . Where the S_i variable is the value of alternative preference, X_{ij} is the variable value of the alternatives on each attribute. The W_j variable is the value of the weight of the criteria, the n variable is the number of criteria, the i variable is an alternative value 1,2,..,m, and the j variable is the criterion value.

$$S_i = \prod_{j=1}^n X_{ij}^{W_j} \quad (13)$$

Step 3: Determine the range of PRTA classification values using the Guttman scale based on Eq. (11). Where the I variable is the interval range, the R variable is the result of the calculation of the highest scores value of S_i minus the lowest score value of S_i , and the K variable is a number of alternatives. The alternative assessment criteria for the PRTA classification are obtained from the result of the calculation of the highest scores value of S_i minus the value of I variable as shown in the result of Eq. (14).

$$\begin{cases} \text{PRTA, if } S_i \geq \text{the scale range} \\ \text{Non_PRTA, if } S_i < \text{the scale range} \end{cases} \quad (14)$$

D. Simple Additive Weighting (SAW)

The SAW method is a multi-process method in spatial decisions making with multicriteria parameters. The SAW method performs a weighted summation of the performance ratings on each alternative attribute. The process of normalizing the decision matrix (X) to a scale that can be compared with all existing alternative ratings [55]. The advantage of the SAW method compared to the decision support system method that involves other multicriteria parameters lies in its ability to make a more precise assessment because it is based on the criteria value and the weight of the level of importance required.

The following sequence of the steps spatial analysis process for the PRTA classification using the SAW method:

Step 1: Follow the steps in the flow in Fig. 2 to define the criteria used as a benchmark for solving the problem and determine the priority of parameter weights.

Step 2: Perform normalization using Eq. 15 for each alternative value on each attribute by calculating the performance rating value.

$$r_{ij} = \begin{cases} \frac{x_{ij}}{\text{Max}_i x_{ij}}, & \text{if } j \text{ is benefit attribute} \\ \frac{\text{Min}_i x_{ij}}{x_{ij}}, & \text{if } j \text{ is cost attribute} \end{cases} \quad (15)$$

where r_{ij} variable is the normalized performance rating of alternative A_i on attributes C_j and j , $\text{Max } X_{ij}$ variable is the greatest value of each criterion i , and $\text{Min } X_{ij}$ variable is the smallest value of each criterion i , X_{ij} variable is the attribute values that each criterion has. If the largest value is the best, then it is included in the benefit attribute category. If the smallest value is the best, then it is included in the cost attribute category.

Step 3: Calculate the value of preference weight on each alternative (V_i) using Eq. (16). Where the V_i variable is the ranking for each alternative. the W_j variable is the ranking weight value of each criterion, and r_{ij} variable is the normalized performance rating value.

$$V_i = \sum_{j=1}^n w_j r_{ij} \quad (16)$$

Step 4: Determine the range of PRTA classification values using the Guttman scale based on Eq. (11). Where the I variable is the interval range, the R variable is the result of the calculation of the highest scores value of V_i minus the lowest

score value of V_i , and the K variable is a number of alternatives. The alternative assessment criteria for the PRTA classification are obtained from the result of the calculation of the highest scores value of V_i minus the value of I variable as shown in the result of Eq. (17).

$$\begin{cases} \text{PRTA, if } V_i \geq \text{the scale range} \\ \text{Non_PRTA, if } V_i < \text{the scale range} \end{cases} \quad (17)$$

E. Weighted Product Model (WPM)

Spatial data modeling with the WPM method is a process to determine the weight of the priority value on each attribute parameter criterion, perform weighting by dividing the attribute weights by the weight of all attributes to get the total value equal to 1, determining the total vector value S to produce the vector V in produce the highest value that will be used as an alternative selection [51]–[53] The WPM method can be used for MCDM single or multi-dimensional categories [56].

The spatial data modeling process for the PRTA classification using the WPM method. The following sequence of the steps is as follows:

Step 1: Follow the steps in the flow in Fig. 2 to define the criteria that will be used as a benchmark for solving the problem and determine the priority of parameter weights.

Step 2: Calculate the normalized decision matrix value using Eq. (18) to get the alternative preference value of each criterion represented by the vector S_i . Where the S_i variable is the value of alternative preference, X_{ij} is the variable value of the alternatives on each attribute, and the W_j variable is the value of the weight of the criteria, the n variable is the number of criteria, the i variable is an alternative value $1, 2, \dots, m$, the j variable is the criterion value.

$$S_i = \prod_{j=1}^n x_{ij}^{w_j}, i = 1, 2, 3, \dots, m \quad (18)$$

The normalized decision matrix value is calculated in order to obtain the x_{ij} value by providing the i -th alternative performance rating value on the j -th sub-criteria in the normalized decision matrix value computation. Furthermore, the value of the performance rating is elevated to the relative weight value (w_j), where w_j will be positive for the benefit attribute and negative for the cost attribute, depending on the attribute being evaluated. The sum of the w_j values for each sub-criteria on the same criteria will be worth 1. The value of w_j is calculated using Eq. 4.19.

$$w_j = \frac{w_j}{\sum w_j} \quad (19)$$

Step 3: Calculate the relative preference value of each alternative V_i using Eq. 20. Where, V_i variable is the relative preference of each i -th alternative. x_{ij} variable is the criteria value for each alternative to the i -th and the criteria j -th. w_j variable is the weight of the criteria or sub-criteria and the n variable is the number of criteria.

$$V_i = \frac{\prod_{j=1}^n x_{ij}^{w_j}}{\prod_{j=1}^n x_j^{*w_j}}, i = 1, 2, 3, \dots, m \quad (20)$$

Step 4: Determine the range of PRTA classification values using the Guttman scale based on Eq. (11). Where the I variable is the interval range, the R variable is the result of the calculation of the highest scores value of V_i minus the lowest score value of V_i , and the K variable is a number of alternatives. The alternative assessment criteria for the PRTA classification are obtained from the result of the calculation of the highest scores value of V_i minus the value of I variable as shown in the result of Eq. (21).

$$\begin{cases} \text{PRTA, if } V_i \geq \text{the scale range} \\ \text{Non_PRTA, if } V_i < \text{the scale range} \end{cases} \quad (21)$$

F. Multi-Attribute Utility Theory (MAUT)

Spatial data modeling with the MAUT method is to determine the value of $U(A_i)$ With the weight value on each sub-criteria parameter and the priority value of each attribute parameter's interest, calculate the number of criteria in each attribute [57][58]. the more value of sub-criterion of every single parameter, the obtained value will end up with a high value $U(A_i)$ [59].

The MAUT method will change from several parameters of importance to a numerical value with a scale of 1-5, where a scale of 1 is the worst choice, and a scale of 5 is the best choice. The results of the MAUT method will provide a ranking order of alternative evaluations that describe the choices of policymakers. The spatial data modeling process for the PRTA classification using the MAUT method. The following sequence of the steps is as follows:

Step 1: Follow the steps in the flow Fig. 2 to define the criteria used as a benchmark for solving the problem and determine the priority of parameter weights.

Step 2: Make the normalized matrix using Eq. (1). Where the $U(x)$ variable is the normalized alternative weight, the x is the alternative weight, the x_i^- is the minimum weight of the x -th criterion, and the x_i^+ is the maximum weight of the x -th criterion using Eq. (22).

$$U_{(x)} = \frac{x - x_i^-}{x_i^+ - x_i^-} \quad (22)$$

Step 3: Calculate the evaluation value of each alternative $V_{(x)}$ by multiplying utility $U_{(x)}$ by weight using Eq. (23).

$$V_{(x)} = \sum_{i=1}^n w_j * x_{ij} \quad (23)$$

Where the $V_{(x)}$ variable is the evaluation value of each alternative of the PRTA classification for the i -th data, the value of the division between the parameter weighting value and the number of sub-criteria on each parameter then multiplied by the weight of the attribute priority value at each parameter criteria. The w_k variable is the weight of the attribute sub-criterion on each parameter of the parameter until the k -th data and $u_k(x_{ik})$ is the parameter of the k -th data multiplied by the priority value of each parameter x_{ik} . The A_i variable is the weighting value of multicriteria parameters.

Step 4: Determine the range of PRTA classification values using the Guttman scale based on Eq. (11). Where the I

variable is the interval range, the R variable is the result of the calculation of the highest scores value of V_x minus the lowest score value of V_x , and the K variable is a number of alternatives. The alternative assessment criteria for the PRTA classification are obtained from the result of the calculation of the highest scores value of V_x minus the value of the I variable as shown in the result of Eq. (24).

$$\begin{cases} \text{PRTA, if } V_x \geq \text{the scale range} \\ \text{Non_PRTA, if } V_x < \text{the scale range} \end{cases} \quad (24)$$

G. Technique for Others Reference by Similarity to Ideal Solution (TOPSIS)

The TOPSIS method is a decision-making method that involves multicriteria parameters used to overcome alternative problems due to uncertainty/inconsistency [60]–[62] The TOPSIS method also determines the distance of the ideal solution to smaller and larger before making the determination of alternative value with the result of alternative calculation has the final value < 1 [50]–[53] The concept of selecting the best alternative in the TOPSIS method is that the best-selected alternative consists of alternatives with the shortest distance from the positive ideal solution and the longest distance from the negative ideal solution.

The spatial data modeling process for the PRTA classification using the MAUT method used the following sequence of the steps is as follows:

Step 1: Follow the steps in the flow in Fig. 2 to define the criteria used as a benchmark for solving the problem and determine the priority of parameter weights.

Step 2: Calculate a normalized decision matrix using Eq. (25) [63]. Where the r_{yx} variable is the normalized value for each y -th alternative to the x -th criteria with $i=1,2,\dots,m$ and $j=1,2,\dots,n$.

$$r_{yx} = \frac{x_{yx}}{\sqrt{\sum_{i=1}^m x_{yx}^2}} \quad (25)$$

Step 3: Calculate a weighted normalized decision matrix using Eq. (26). Multiply the weight of the parameter criteria with the value of each attribute.

$$v_{yx} = W_{yx} * r_{yx} \quad (26)$$

Where v_{yx} variable is the weighted normalized value, the variable w_{yx} is the weight of each criterion, and the variable r_{yx} is the normalized value of each alternative against the j -th criterion with $i=1,2,\dots,m$ and $j=1,2,\dots,n$.

Step 4: Calculate the ideal solution based on the maximum value of A^+ using Eq. (27) [51] [64] and the negative ideal solution based on a minimum value of A^- using Eq. (28) [51] [64].

$$A_x^+ = \{v_{y1}^+, v_{y2}^+, v_{y3}^+\}$$

$$A^- = \{\max y_1, \max v_2, \max v_3, \dots, \max v_n\}$$

Where,

$$v_x^+ = \begin{cases} \max v_{yx}; & \text{if } yx \text{ is benefit attribute} \\ \min v_{yx}; & \text{if } yx \text{ is cost attribute} \end{cases} \quad (27)$$

$$A_x^- = \{v_{y1}^-, v_{y2}^-, v_{y3}^-\}$$

$$A^- = \{\min v_1, \min v_2, \min v_3, \dots, \min v_n\}$$

Where,

$$v_x^- = \begin{cases} \max v_{yx}; & \text{if } yx \text{ is benefit attribute} \\ \min v_{yx}; & \text{if } yx \text{ is cost attribute} \end{cases} \quad (28)$$

Step 5: Calculate the positive and negative ideal solution spacing, as referenced in Eq.s (29) and (30) [51] [64]. In this research, the ideal positive solution is based on the maximum value of D + of the Eq. (29) [51] [64] and the ideal negative solution distance based on a minimum value of D - of Eq. (30) [51] [64].

$$D_y^+ = \sqrt{\sum_{y=1}^n (v_{yx} - A_x^+)^2} \quad (29)$$

$$D_y^- = \sqrt{\sum_{y=1}^n (v_{yx} - A_x^-)^2} \quad (30)$$

Where the D_y^+ variable is used to calculate the maximum ideal solution distance as much as the y-th data. The D_y^- variable is used to calculate the minimum ideal solution distance as much as the y-th data.

Step 6: Calculate the preference value for each alternative to be generated by Eq. (31) [51] [64].

$$V = \frac{D_i^-}{D_i^- + D_i^+} \quad (31)$$

where D_i^- is the ideal minimal solution distance value of the i-th data dan D_i^+ is the maximum ideal solution distance value as much as a number of i data.

Step 7: Determine the range of PRTA classification values using the Guttman scale based on Eq. (11). Where the I variable is the interval range, the R variable is the result of the calculation of the highest scores value of V_i minus the lowest score value of V_i , and the K variable is a number of alternatives. The alternative assessment criteria for the PRTA classification are obtained from the result of the calculation of the highest scores value of V_i minus the value of I variable as shown in Eq. (32) .

$$\begin{cases} \text{PRTA, if } V_i \geq \text{the scale range} \\ \text{Non_PRTA, if } V_i < \text{the scale range} \end{cases} \quad (32)$$

H. Analytical Hierarchy Process (AHP)

The Analytical Hierarchy Process (AHP) is a pairwise comparison method through an analytic hierarchy process, where the parameter weights are determined by normalization through the eigenvectors associated with the maximum eigenvalues in the unit ratio matrix. The weighting between parameters in this study is accomplished by the AHP method approach based on flow depicted in Fig. 2. Which consists of the following:

Step 1-9: Use the process in section III subsection A for the priority weight of the parameters using the AHP method.

Step 10: Determine the range of PRTA classification values using the Guttman scale based on Eq. (11). Where the I

variable is the interval range, the R variable is the result of the calculation of the highest scores value of CR variable minus the lowest score value of CR variable, and K variable is a number of alternatives. The alternative assessment criteria for the PRTA classification are obtained from the result of the calculation of the highest scores value of CR variable minus the value of I variable, as shown in the result of Eq. (33).

$$\begin{cases} \text{PRTA, if } CR \geq \text{the scale range} \\ \text{Non_PRTA, if } CR < \text{the scale range} \end{cases} \quad (33)$$

I. Multicriteria Evaluation techniques (MCE)

This research method evaluation uses the accuracy, and F1 score approaches. The F1 score is obtained from the values of precision and recall. A confusion matrix [65] is used in this evaluation technique, consisting of two positive classes and a negative class to compare actual data and classification data [66]. Multi-class classification [65] is used in the discussion of this paper: prone road traffic accident (PRTA), and non-prone road traffic accident (Non-PRTA). The precision and recall value is calculated with the average value in each class.

Accuracy in the measurement of a method is used to determine the accuracy value in clarifying the results of classification data with actual data with Eq. (34) [65]. Precision describes the amount of positive-valued data divided by total positive-valued data in Eq. (35) [65]. The recall describes the percentage of data in the positive category classified by the system with the calculation in Eq. (36) [65]. Results of precision and recall values are used to calculate F1-score, as in formula (37) [65]. The accuracy of the data generated in classification is known from the percentage after testing between the actual data in the form of an analog map of the classification of the watershed erosion zone and prediction data with MAUT, WPM, WSM, and TOPSIS methods. Performance value classification with categories 91% – 100% is very good classification, 81% – 90% is good classification, 71% – 80% is fair classification, 61% – 70% is poor classification, and values below 60% are false classification [67].

$$Accuracy_M = \frac{\sum_{i=1}^l \frac{TP_i + TN_i}{TP_i + FN_i + FP_i + TN_i}}{l} \quad (34)$$

$$Precision_M = \frac{\sum_{i=1}^l \frac{TP_i}{TP_i + FP_i}}{l} \quad (35)$$

$$Recall_M = \frac{\sum_{i=1}^l \frac{TP_i}{TP_i + FN_i}}{l} \quad (36)$$

$$F - score_M = \frac{(\beta^2 + 1) Precision_M Recall_M}{\beta^2 Precision_M + Recall_M} \quad (37)$$

Where TP_i is the amount of data + which when the classification is true by the method used for the i-th class. TN_i is the amount of data. When the classification is true by the method used for the i-th class. FP_i is the amount of + data that is classified as false by the method used for the i-th class. FN_i is the amount of data - which when the classification is false by the method used for the i-th class. l is the number of classification classes. Average accuracy is the average value of method accuracy in all classification class. $Precision_\mu$ is the precision value of each classification class. $Precision_M$

represents the average value of the precision in all classification classes. $Recall_{\mu}$ is the recall value of each classification class. $Recall_M$ represents the average value of recalls in all classification classes. $Fscore_{\mu}$ is a performance matrix to calculate the average of precision and recall values in each classification class. $F-score_M$ is a performance matrix to calculate the average of precision and recall values in all classification classes.

IV. RESULT AND DISCUSSION

Based on the private spatial datasets and quantitative attribute data explained in section III, the results of this study are discussed in the following subsections.

A. Parameter Priority Weight

In each of the methods used in the MCDM, the weight of each parameter priority value in this study uses the opinion of EJ (score) and mathematical calculations using AHP based on the score given by EJ (EJ-AHP). Tables IV and V are the results of mathematical calculations of the pairwise comparison matrix of the AHP method to produce the priority weight of the parameters based on the flow in Fig. 2 with the process in Section III for sub-section A.

TABLE III. PARAMETER PRIORITY VALUE WEIGHTING RESULTS

Parameters Symbol	AHP Weight
VCR (X1)	0.02
IRI (X2)	0.06
HA (X3)	0.10
VA (X4)	0.14
Vr (X5)	0.18
Road Type (X6)	0.22
Shoulder (X7)	0.27
Total:	1.00

TABLE IV. SUB-PARAMETER PRIORITY VALUE WEIGHTING RESULTS

Parameters	Multicriteria Parameters	EJ Scoring	EJ-AHP Weight
VCR (%)	$VCR \geq 0.85$ && $VCR < 1.00$	5	0.34
	$VCR \geq 0.70$ && $VCR < 0.85$	4	0.26
	$VCR \geq 0.45$ && $VCR < 0.70$	3	0.24
	$VCR \geq 0.20$ && $VCR < 0.45$	2	0.12
	$VCR < 0.20$	1	0.04
IRI (m/km)	$IRI \geq 12$	4	0.43
	$IRI \geq 8$ && $IRI < 12$	3	0.35
	$IRI \geq 4$ && $IRI < 8$	2	0.17
	$IRI < 4$	1	0.05
HA (rad/km)	$HA \geq 3.50$	3	0.54
	$HA \geq 0.25$ && $HA < 3.50$	2	0.37
	$HA < 0.25$	1	0.09
VA (m/km)	$VA \geq 45$	3	0.54
	$VA \geq 5$ && $VA < 45$	2	0.37
	$VA < 5$	1	0.09

Vr (km/jam)	$Vr \geq 100$	6	0.33
	$Vr \geq 80$ && $Vr < 100$	5	0.25
	$Vr \geq 65$ && $Vr < 80$	4	0.17
	$Vr \geq 60$ && $Vr < 65$	3	0.16
	$Vr \geq 50$ && $Vr < 60$	2	0.06
	$Vr < 50$	1	0.03
Road Type	2/2 UD	5	0.36
	4/2 UD	4	0.29
	4/2 D	3	0.24
	6/2 D	2	0.08
	2/1 UD	1	0.04
Shoulder	No	2	0.83
	Yes	1	0.17
Parameters	Range	EJ Scoring	EJ-AHP Weight
VCR (%)	$VCR \geq 0.90$ && $VCR < 1.00$	5	0.34
	$VCR \geq 0.75$ && $VCR < 0.90$	4	0.26
	$VCR \geq 0.50$ && $VCR < 0.75$	3	0.24
	$VCR \geq 0.30$ && $VCR < 0.50$	2	0.12
	$VCR < 0.30$	1	0.04
IRI (m/km)	$IRI \geq 12$	4	0.43
	$IRI \geq 8$ && $IRI < 12$	3	0.35
	$IRI \geq 4$ && $IRI < 8$	2	0.17
	$IRI < 4$	1	0.05
HA (rad/km)	$HA \geq 3.50$	3	0.54
	$HA \geq 0.25$ && $HA < 3.50$	2	0.37
	$HA < 0.25$	1	0.09
VA (m/km)	$VA \geq 45$	3	0.54
	$VA \geq 5$ && $VA < 45$	2	0.37
	$VA < 5$	1	0.09
Vr (km/jam)	$Vr \geq 100$	6	0.33
	$Vr \geq 90$ && $Vr < 100$	5	0.25
	$Vr \geq 75$ && $Vr < 90$	4	0.17
	$Vr \geq 60$ && $Vr < 75$	3	0.16
	$Vr \geq 50$ && $Vr < 60$	2	0.06
	$Vr < 50$	1	0.03
Road Type	2/2 UD	5	0.36
	4/2 UD	4	0.29
	4/2 D	3	0.24
	6/2 D	2	0.08
	2/1 UD	1	0.04
Shoulder	No	2	0.83
	Yes	1	0.17

B. The Guttman Scale to Determine the Classification of Accident Prone Roads

The Guttman scale [68] is used to measure the generated classification values in this paper. This scale is used to draw conclusions from qualitative data [69]. It is also used to estimate the value of the classification resulting in an intervention value that is still ambiguous due to uncertainty [70]. It is possible to assess the uncertainty factor of a variable class defined using the Guttman scale [71] in Eq. (11) for a dataset that employs a weight in the analysis process and delivers a value.

The test data consisted of 180 primary arterial roads and 201 primary collector roads, where the data is categorized as a small-scale dataset. The value of the scale on the SAW, WP, SAW, WPM, MAUT, TOPSIS, and AHP methods using Eq. (12), (14), (17), (21), (24), (32), and (33) based on the process calculations in section III sub-sections B to H, respectively.

TABLE V. THE SCALE TO DETERMINE THE CLASSIFICATION OF ACCIDENT PRONE ROADS

MCDM Models	Arterial Road Scale	Collector Road Scale
WSM	$\begin{cases} PRTA, \text{if } A_i \geq 0,2729 \\ Non_PRTA, \text{if } A_i < 0,2729 \end{cases}$	$\begin{cases} PRTA, \text{if } A_i \geq 0,2886 \\ Non_PRTA, \text{if } A_i < 0,2886 \end{cases}$
WP	$\begin{cases} PRTA, \text{if } S_i \geq 0,2367 \\ Non_PRTA, \text{if } S_i < 0,2367 \end{cases}$	$\begin{cases} PRTA, \text{if } S_i \geq 0,2579 \\ Non_PRTA, \text{if } S_i < 0,2579 \end{cases}$
SAW	$\begin{cases} PRTA, \text{if } V_i \geq 0,5753 \\ Non_PRTA, \text{if } V_i < 0,5753 \end{cases}$	$\begin{cases} PRTA, \text{if } V_i \geq 0,6037 \\ Non_PRTA, \text{if } V_i < 0,6037 \end{cases}$
WPM	$\begin{cases} PRTA, \text{if } V_i \geq 0,0060 \\ Non_PRTA, \text{if } V_i < 0,0060 \end{cases}$	$\begin{cases} PRTA, \text{if } V_i \geq 0,0055 \\ Non_PRTA, \text{if } V_i < 0,0055 \end{cases}$
MAUT	$\begin{cases} PRTA, \text{if } V_x \geq 0,2886 \\ Non_PRTA, \text{if } V_x < 0,2886 \end{cases}$	$\begin{cases} PRTA, \text{if } V_x \geq 0,4723 \\ Non_PRTA, \text{if } V_x < 0,4723 \end{cases}$
TOPSIS	$\begin{cases} PRTA, \text{if } V_i \geq 0,4073 \\ Non_PRTA, \text{if } V_i < 0,4073 \end{cases}$	$\begin{cases} PRTA, \text{if } V_i \geq 0,4157 \\ Non_PRTA, \text{if } V_i < 0,4157 \end{cases}$
AHP	$\begin{cases} PRTA, \text{if } CR \geq 0,00229 \\ Non_PRTA, \text{if } CR < 0,00229 \end{cases}$	$\begin{cases} PRTA, \text{if } CR \geq 0,000080 \\ Non_PRTA, \text{if } CR < 0,000080 \end{cases}$

C. Model Performance Evaluation

The MCDM spatial analysis model was developed to assist the decision-making process by selecting alternatives in the multi-class classification [52][72]. The steps in the MCDM model are to determine the multicriteria parameter that will be an alternative to the multi-class classification, to describe the quantitative data requirements that will be processed to have an impact on the alternatives of the multi-class being processed, then to process the numerical values on the

qualitative data to determine the rating on each of the multicriteria parameters.

Table VI results from multicriteria evaluation techniques using a confusion matrix based on the process in section III sub-section I. The accuracy values in the experimental test of arterial road type data using the WSM and TOPSIS methods were 63% superior to other methods, followed by the MAUT, SAW, WP, and WPM methods, and the AHP method, namely 59%, 58%, 54%, 43%, respectively. However, the AHP method is superior in the collector road type experiment with an accuracy value of 70%, followed by the TOPSIS and WSM, SAW, MAUT, WP, and WPM methods, namely 58%, 57%, 53%, 41%, respectively.

Fig. 3 is a sampling test of the spatial analysis results of accident-prone road classification using the WSM method on the North Rim Probolinggo arterial road type, Indonesia. Calculate the weight value for each multicriteria parameter using Flow in Fig. 2 with the results in Tables IV and V. Perform the calculation process based on section III subsection B, then obtain the value of the variable A_i is 0.3593, referring to Table VI, the road is included in the PRTA classification.

TABLE VI. THE MCDM MODEL PERFORMANCE

MCDM Models	Accuracy	F1-Score
Arterial Roads		
WSM	63%	54%
WP	54%	55%
SAW	58%	52%
WPM	54%	55%
MAUT	59%	53%
TOPSIS	63%	54%
AHP	43%	53%
Collector Roads		
WSM	58%	49%
WP	41%	45%
SAW	57%	48%
WPM	41%	45%
MAUT	53%	47%
TOPSIS	58%	49%
AHP	70%	47%

- J. Telecommun. Electron. Comput. Eng., vol. 10, no. 2–3, pp. 83–89, 2018.
- [17] A. V. Vitaningsih and D. Cahyono, “Geographical Information System for Mapping Road Using Multi-Attribute Utility Method,” in International Conference on Science and Technology-Computer (ICST), 2016, pp. 0–4.
- [18] A. V. Vitaningsih and D. Cahyono, “Geographical Information System for mapping accident-prone roads and development of new road using Multi-Attribute Utility method,” in Proceedings - 2016 2nd International Conference on Science and Technology-Computer, ICST 2016, 2017, pp. 66–70.
- [19] N. Mahmoody Vanolya, M. Jelokhani-Niaraki, and A. Toomanian, “Validation of spatial multicriteria decision analysis results using public participation GIS,” *Appl. Geogr.*, vol. 112, no. November, pp. 1–16, 2019.
- [20] W. Alkhadour, J. Zraqou, A. Al-Helali, and S. Al-Ghananeem, “Traffic Accidents Detection using Geographic Information Systems (GIS): Spatial Correlation of Traffic Accidents in the City of Amman, Jordan,” *Int. J. Adv. Comput. Sci. Appl.*, vol. 12, no. 4, pp. 484–494, 2021.
- [21] S. G. and N. D. Bekir Çetintav, Gözde Ulutagay, “A New Approach Of Combining Expert Judgment And Data Knowledge In Multi-Attribute Decision Making,” in Uncertainty Modelling in Knowledge Engineering and Decision Making, 2016, pp. 112–118.
- [22] T. Chen, C. Zhang, and L. Xu, “Factor analysis of fatal road traffic crashes with massive casualties in,” *Adv. Mech. Eng.*, vol. 8, no. 4, pp. 1–11, 2016.
- [23] A. C. L. Vieira, M. D. Oliveira, and C. A. Bana e Costa, “Enhancing knowledge construction processes within multicriteria decision analysis: The Collaborative Value Modelling framework,” *Omega (United Kingdom)*, vol. 94, no. July, pp. 1–36, 2020.
- [24] H. Fang and Z. Guo, “Vessel Collision Accidents Analysis based on Factor Analysis and GA-SVM *,” in IEEE 2nd Advanced Information Technology, Electronic and Automation Control Conference (IAEAC), 2017, pp. 191–195.
- [25] E. Turunen, “Using GUHA Data Mining Method in Analyzing Road Traffic Accidents Occurred in the Years 2004–2008 in Finland,” *Data Sci. Eng.*, vol. 2, no. 3, pp. 224–231, 2017.
- [26] X. Qu, W. Wang, W. fu Wang, and P. Liu, “Real-time rear-end crash potential prediction on freeways,” *J. Cent. South Univ.*, vol. 24, no. 11, pp. 2664–2673, 2017.
- [27] M. Juhász and C. Koren, “Getting an Insight into the Effects of Traffic Calming Measures on Road Safety,” *Transp. Res. Procedia*, vol. 14, pp. 3811–3820, 2016.
- [28] D. Nilsson, M. Lindman, T. Victor, and M. Dozza, “Definition of run-off-road crash clusters—For safety benefit estimation and driver assistance development,” *Accid. Anal. Prev.*, vol. 113, no. November 2017, pp. 97–105, 2018.
- [29] M. Bassani, L. Rossetti, and L. Catani, “Spatial analysis of road crashes involving vulnerable road users in support of road safety management strategies,” in Transportation Research Procedia, 2020, vol. 45, pp. 394–401.
- [30] H. Zhang, X. Wang, J. Cao, M. Tang, and Y. Guo, “A multivariate short-term traffic flow forecasting method based on wavelet analysis and seasonal time series,” *Appl. Intell.*, vol. 48, no. 10, pp. 3827–3838, 2018.
- [31] R. Al-Ruzouq, K. Hamad, S. Abu Dabous, W. Zeiada, M. A. Khalil, and T. Voigt, “Weighted Multi-attribute Framework to Identify Freeway Incident Hot Spots in a Spatiotemporal Context,” *Arab. J. Sci. Eng.*, vol. 44, no. 10, pp. 8205–8223, 2019.
- [32] B. Pradhan and M. Ibrahim Sameen, “Review of Traffic Accident Predictions with Neural Networks,” in Urbanization and Its Impact in Contemporary China, Springer, Cham, 2020, pp. 97–109.
- [33] R. Goel, “Modelling of road traffic fatalities in India,” *Accid. Anal. Prev.*, vol. 112, no. October, pp. 105–115, 2018.
- [34] M. Keymanesh, H. Ziari, S. Roudini, and A. N. Ahangar, “Identification and Prioritization of ‘Black Spots’ without Using Accident Information,” *Model. Simul. Eng.*, vol. 2017, 2017.
- [35] Ö. Kaya, A. Tortum, K. D. Alemdar, and M. Y. Çodur, “Site selection for EVCS in Istanbul by GIS and multi-criteria decision-making,” *Transp. Res. Part D Transp. Environ.*, vol. 80, no. February, p. 102271, 2020.
- [36] M. S. Fatemeh Haghghat, “Application of a Multi-Criteria Approach To Road Safety Evaluation in the Bushehr Province , Iran,” *Traffic Plan. Prelim. Commun.*, vol. 23, no. 5, pp. 341–352, 2011.
- [37] S. Da Costa, X. Qu, and P. M. Parajuli, “A Crash Severity-Based Black Spot Identification Model,” *J. Transp. Saf. Secur.*, vol. 7, no. 3, pp. 268–277, 2015.
- [38] F. Torrieri and A. Batà, “Spatial multi-Criteria decision support system and strategic environmental assessment: A case study,” *Buildings*, vol. 7, no. 4, 2017.
- [39] S. Kumar and D. Toshniwal, “A data mining approach to characterize road accident locations,” *J. Mod. Transp.*, vol. 24, no. 1, pp. 62–72, 2016.
- [40] T. Sipos, “Spatial statistical analysis of the traffic accidents,” *Period. Polytech. Transp. Eng.*, vol. 45, no. 2, pp. 101–105, 2017.
- [41] R. Mosadeghi, J. Warnken, R. Tomlinson, and H. Mirfenderesk, “Comparison of Fuzzy-AHP and AHP in a spatial multi-criteria decision making model for urban land-use planning,” *Comput. Environ. Urban Syst.*, vol. 49, pp. 54–65, 2015.
- [42] F. Yakar, “A multicriteria decision making-based methodology to identify accident-prone road sections,” *J. Transp. Saf. Secur.*, pp. 1–15, 2019.
- [43] A. G. Macbeth and R. Eng, “Road Classification Systems – Christchurch and Toronto,” in 2001 Traffic Management Workshop Auckland, 2001, no. 03, pp. 1–12.
- [44] E. M. Setton, P. W. Hystad, and C. P. Keller, “Road Classification Schemes – Good Indicators of Traffic Volume?,” in Spatial Sciences Laboratories Occasional Papers, 2005, vol. i, pp. 1–11.
- [45] A. Suraji, L. Djakfar, and A. Wicaksono, “Analysis of bus performance on the risk of traffic accidents in East Java-Indonesia,” *EUREKA, Phys. Eng.*, vol. 2021, no. 3, pp. 111–118, 2021.
- [46] H. Pilko, S. Mandžuka, and D. Barić, “Urban single-lane roundabouts: A new analytical approach using multi-criteria and simultaneous multi-objective optimization of geometry design, efficiency and safety,” *Transp. Res. Part C Emerg. Technol.*, vol. 80, no. July, pp. 257–271, 2017.
- [47] M. Rosić, D. Pešić, D. Kukić, B. Antić, and M. Božović, “Method for selection of optimal road safety composite index with examples from DEA and TOPSIS method,” *Accid. Anal. Prev.*, vol. 98, no. January, pp. 277–286, 2017.
- [48] A. Ait-Mlouk, T. Agouti, and F. Gharnati, “Mining and prioritization of association rules for big data: multi-criteria decision analysis approach,” *J. Big Data*, vol. 4, no. 1, pp. 1–21, 2017.
- [49] T. L. Saaty, “A scaling method for priorities in hierarchical structures,” *J. Math. Psychol.*, vol. 15, no. 3, pp. 234–281, 1977.
- [50] E. Triantaphyllou and S. H. Mann, “An examination of the effectiveness of multi-dimensional decision-making methods: A decision-making paradox,” *Decis. Support Syst.*, vol. 5, no. 3, pp. 303–312, 1989.
- [51] E. Triantaphyllou, *Multi-Criteria Decision Making Methods: A Comparative Study*. Springer, Boston, MA, 2000.
- [52] E. Mulliner, N. Malys, and V. Maliene, “Comparative analysis of MCDM methods for the assessment of sustainable housing affordability,” *Omega (United Kingdom)*, vol. 59, pp. 146–156, 2016.
- [53] V. Maliene, R. Dixon-Gough, and N. Malys, “Dispersion of relative importance values contributes to the ranking uncertainty: Sensitivity analysis of Multiple Criteria Decision-Making methods,” *Appl. Soft Comput. J.*, vol. 67, pp. 286–298, 2018.
- [54] C. H. Yeh, “A Problem-based Selection of Multi-attribute Decision-making Methods,” *Int. Trans. Oper. Res.*, vol. 9, no. 2, pp. 169–181, 2002.
- [55] E. Boltürk, A. Karaşan, and C. Kahraman, “Simple additive weighting and weighted product methods using neutrosophic sets,” in *Studies in Fuzziness and Soft Computing*, vol. 369, 2019, pp. 647–676.

- [56] Z. Chourabi, F. Khedher, A. Babay, and M. Cheikhrouhou, "Multi-criteria decision making in workforce choice using AHP, WSM and WPM," *J. Text. Inst.*, vol. 110, no. 7, pp. 1092–1101, 2019.
- [57] J. S. Dyer, "MAUT — Multiattribute Utility Theory," in *Multiple Criteria Decision Analysis: State of the Art Surveys*, 2005, pp. 265–292.
- [58] M. Bystrzanowska and M. Tobiszewski, "How can analysts use multicriteria decision analysis?," *TrAC - Trends Anal. Chem.*, vol. 105, pp. 98–105, 2018.
- [59] M. Cinelli, S. R. Coles, and K. Kirwan, "Analysis of the potentials of multi criteria decision analysis methods to conduct sustainability assessment," *Ecol. Indic.*, vol. 46, pp. 138–148, 2014.
- [60] K. Mela, T. Tiainen, and M. Heinisuo, "Comparative study of multiple criteria decision making methods for building design," *Adv. Eng. Informatics*, vol. 26, no. 4, pp. 716–726, 2012.
- [61] S. Başaran and F. El Homsy, "Mobile Mathematics Learning Application Selection using Fuzzy TOPSIS," *Int. J. Adv. Comput. Sci. Appl.*, vol. 13, no. 2, pp. 270–282, 2022.
- [62] N. A. M. Zulkefli, M. Madanan, T. M. Hardan, and M. H. M. Adnan, "Multi-Criteria Prediction Framework for the Prioritization of Council Candidates based on Integrated AHP-Consensus and TOPSIS Methods," *Int. J. Adv. Comput. Sci. Appl.*, vol. 13, no. 2, pp. 352–359, 2022.
- [63] L. Vargas, *Why the Analytic Hierarchy Process Is Not Like Multiattribute Utility Theory*, vol. 356, 1989.
- [64] T. Vulevic, N. Dragovic, S. Kostadinov, S. Belanovic Simic, and I. Milovanovic, "Prioritization of Soil Erosion Vulnerable Areas Using Multi-Criteria Analysis Methods," *Polish J. Environ. Stud.*, vol. 24, no. 1, pp. 317–323, 2015.
- [65] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," *Inf. Process. Manag.*, vol. 45, no. 4, pp. 427–437, 2009.
- [66] Max Bramer, *Principles Data Mining*. London-Spinger: <http://www.springer.com/series/7592>, 2007.
- [67] Florin Gorunescu, *Data Mining: Concept, Models, Techniques*. Springer, 2011.
- [68] L. Guttman, "The Determinacy Of Factor Score Matrices With Implications For Five Other Basic Problems Of Common-Factor Theory," *Br. J. Stat. Psychol.*, vol. 7, no. 2, pp. 65–81, 1955.
- [69] L. Guttman, "A Basis for Scaling Qualitative Data," *Am. Sociol. Rev.*, vol. 9, no. 2, p. 139, 1944.
- [70] R. E. Tractenberg, F. Yumoto, P. S. Aisen, J. A. Kaye, and R. J. Mislevy, "Using the Guttman scale to define and estimate measurement error in items over time: The case of cognitive decline and the meaning of 'points lost,'" *PLoS One*, vol. 7, no. 2, pp. 1–8, 2012.
- [71] A. Stegeman, "A new method for simultaneous estimation of the factor model parameters, factor scores, and unique parts," *Comput. Stat. Data Anal.*, vol. 99, no. July, pp. 189–203, 2016.
- [72] M. R. Asadabadi, "The stratified multi-criteria decision-making method," *Knowledge-Based Syst.*, vol. 162, no. December, pp. 115–123, 2018.

Application of the Fuzzy Delphi Method to Identify and Prioritize the Social-Health Family Disintegration Indicators in Yemen

Abed Saif Ahmed Alghawli^{1*}

Department of Computer Science, College of Sciences and Humanities
Prince Sattam Bin Abdulaziz University, Aflaj
Kingdom of Saudi Arabia

Abdualmajed A. Al-khulaidi²

Department of Computer Science
Faculty of Computer and Information Technology
Sana'a University, Sana'a, Yemen

Nesmah A. AL-Khulaidi⁴

Department of Computer Networks
Yemeni Academy for Graduate Studies
Sana'a, Yemen

Adel A. Nasser³

Department of Information Systems, Faculty of Science
Sa'adah University, Sa'adah, Yemen & Modern Specialized
College of Medical and Technical Sciences, Sana'a, Yemen

Faisal A. Abass⁵

Yemen Center of Social Studies &
Labor Research, Sana'a University
Sana'a, Yemen

Abstract—Constantly increasing political events and socially related changes have led governments worldwide to adopt strategies to reduce their negative effects on the cohesion of societies, which requires developing assessment frameworks that include realistic, measurable, and useful indicators for analyzing the family disintegration causes, taking into consideration the circumstances surrounding the countries and the development trends that they adopt. Therefore, this study aims to identify and prioritize indicators of a decision-making support framework for evaluation, ranking, and structural comparison of the family disintegration causes resulting from the child marriage phenomenon in Yemen. To achieve this, the Fuzzy Delphi Method was applied. Firstly, a set of related literature and theories were analyzed to extract the expected framework's suitable initial indicators. Then, with the participation of twenty-four local experts, the extracted factors were revised, and the most suitable factors were selected. As a result, one social factor out of nine social-health factors was excluded due to its inappropriateness, and a framework of eight indicators was built. Also, with a rating average of 0.727, it was consistently agreed that the indicator "Increasing divorce rates in marital cases that do not take place according to the common desire of the spouses" is the most important indicator. Also, with high consistent evaluation averages (0.652–0.658), all three health indicators were ranked in the second and third places, while the other four social indicators were ranked in the last three positions (fourth–sixth). Finally, the real applications of the proposed framework were recommended.

Keywords—Family disintegration; child marriage; early marriage; Fuzzy Delphi Method; multi-criteria decision making; Yemen

I. INTRODUCTION

The family system is considered one of the oldest social systems throughout history, and it plays an essential role in the raising and upbringing of children and in determining their orientations, tendencies, and desires within the framework of customs and traditions that link the members of each system together. At the same time, individuals of the same society's family systems are linked by numerous cohesive linkages and relationships that contribute to the creation and growth of socially cohesive societies and generations capable of working, giving, and creating. And in return, the disintegration of family bonds in any society prompts the emergence and development of many types of manifestations that hinder human energies from fulfilling their intended roles and contribute to pushing members to embrace negative behaviors towards their society, which could negatively impact the process of society's development [1]. Therefore, the family is an important part of the social system, and its cohesion has a significant impact on society's progress.

On the other hand, the family is also affected by the society to which it belongs, and its structures, practices, and events fluctuate with changing economic, social, and political conditions and events in it. Such transformations in today's world have a particular impact on the central functions of the family and can sometimes amount to a breach of its social role, which in turn contributes to an increase in antisocial behavior among juveniles, and the intentional or unintentional violations of their civil rights [2]. The persistence of such manifestations in societies often has negative consequences

*Corresponding Author.
Submission Date: April 24, 2022
Acceptance Date: May 12, 2022

and creates new barriers to achieving sustainable development in countries [3], particularly those with high rates of instability, economic hardships, conflicts, and wars, such as Yemen.

Within this context, during the last ten years, the Republic of Yemen has witnessed a lot of political events that eventually led to the deterioration of the sustainable development situation in Yemen, economically, socially, and environmentally [4]. Issues such as poverty, interruption of employee salaries in most government sectors, high and doubling prices of basic goods and services, the displacement of some families from areas experiencing conflicts and wars to relatively safe areas, and the loss of some families for their dependents, with women and children bearing the responsibilities and the brunt of the suffering, are but a few of the manifestations that Yemeni society is experiencing as a result of those events.

As a result of this tragic state of poverty, as well as other factors such as misunderstandings of religion, customs, and traditions, and a lack of awareness, the rate of early marriage for young girls has risen dramatically as one of the negative adaptive methods for survival, with the percentage of girls between the ages of ten and nineteen years marrying accounting for one-eighth of all girls' marriages. In any case, the early marriage phenomenon—as a negative phenomenon imposed on Yemeni society—has become a threat to the local family stability state. In this regard, studies [5] confirmed that it could negatively impact the compatibility between spouses, which constitutes the main element of family stability, not to mention that it is one of the main contributing factors to the high rates of conflict, negative conflict, violence, divorce, psychological cases, and the early death of underage wives [6], [7], [8], [9], which constitute the main causes of family disintegration at the local and international levels [4], [10], [11].

Research problem: Over the past decade, Yemen has been subjected to many political events, economic and cultural changes, causing an increase in the spread of the early marriage phenomenon, which has led to the emergence of risks and threats to the family stability of Yemeni society. This makes the local government responsible for taking objective and effective decisions that contribute to limiting the negative effects of this phenomenon on the life, welfare, and stability of Yemeni society by studying and analyzing the causes of family disintegration related to this phenomenon, prioritizing their treatments according to society's developmental circumstances and its development priorities, and conducting structural comparisons of these causes in different regions, in a manner that ensures optimum utilization and equitable distribution of available and limited resources and budgets. This cannot be practically realized without developing appropriate, realistic, and measurable local indicators through the active participation of a group of local experts.

Accordingly, this study aims to identify and arrange indicators of a local decision-making support framework evaluation, ranking, and structural comparison of the causes of

family disintegration resulting from the phenomenon of early marriage in the Republic of Yemen and through the participation of a group of Yemeni experts.

The remainder of this work is arranged in the following manner. The second section gives a quick summary of relevant sociological theories, their perspectives on family disintegration, and the causes and variables most directly linked to family breakup as a result of early marriage. The concepts and principles of Delphi techniques to enable multi-criteria group decision-making, as well as the proposed Fuzzy Delphi model for solving the problem, were discussed in the third section. In Section 4, the proposed assessment model, and the findings are presented, followed by a discussion, limitations of study, its applications, and future work, and conclusion in the last three parts.

II. RELATED WORK AND LITERATURE REVIEW

This section reviews the literature on the study's topic, including a basic background on the concept of early marriage, the concept and manifestations of family disintegration, key social theories about family disintegration, the factors of family disintegration most associated with the phenomenon of early marriage, and the MCDM tools chosen to solve the problem.

A. The Concept and Causes of Early Marriage in Yemen

The phenomenon of early marriage is particularly prevalent in Yemeni society, especially in rural areas. The causes of this phenomenon have been many. In economic terms, early marriage is an important economic factor for rural families. Yemeni rural families tend to marry their male children at an early age in order to join their children's wives in assisting them by doing the animal and agricultural work they practice as their main source of income. Let them have children early, and also contribute in the future to the same tasks. From another economic perspective, some Yemeni families agree to marry girls at an early age because they are unable to support them for reasons such as poverty, low family income, the death of one or both parents, or other reasons reviewed in the introduction. From a social and cultural standpoint, customs and traditions also play a role in the spread of this negative phenomenon. Males and females are married early for various reasons, such as protecting them from falling into vice, being subjected to violence or harassment, or preventing the spread of extrajudicial sexual relations, which is not accepted by some local cultures. In any case, local studies such as [4], and [7] provide more details about this phenomenon and can be consulted for more information.

From a procedural standpoint, this study defines early marriage as a legitimate marital relationship that begins at a young age for both sexes, or for one of them, and qualifies each other for self-reliance in respect of their obligations to the other, as well as their qualification to have and raise legitimate children born as a result of this relationship, with the early age of that relationship being determined for children under nineteen years of age.

B. The Concept of Family Disintegration and its Manifestations

The literature has reviewed a variety of definitions of family disintegration, which is defined as a situation that occurs as a result of the death of one or both parents [12], divorce [13], the abandonment of the family's head, polygamy, or the absence of the family's head for an extended period of time. Others refer to it as "family dispersion," which can occur as a result of polygamy, the death of one or both parents, or divorce [14]. It also includes residences that have been demolished as a result of a divorce, separation, or the death of one or both parents [15]. Family disintegration is a term used to describe the tension, rupture, or loss of control that occurs within a family system. The collapse of familial bonds and ties between family members, whether by divorce, abandonment, separation, or the loss of one or both parents, whether by death, jail, or otherwise due to particular socio-economic conditions [14], has the same meaning. The disruption of family functions, the collapse of family roles and structure as a result of early marriage, the resulting tension and ongoing family conflicts, or the weakness of family relations are the procedural definitions of family disintegration in this study.

C. Social Theories and Family Disintegration

Based on current and previous studies, the theoretical trends serve as a point of reference for determining the nature of the direction in which the contemporary family has evolved as a result of the social changes that have occurred. Previous family disintegration studies used a variety of theories to investigate the causes and effects of family breakup. This section examines family disintegration studies in the literature and discusses the implications of a number of significant theories, including functional constructivism, conflict theory, and symbolic interaction theory.

1) *Functional constructive theory*: According to this idea, family stability is accomplished by completing the social functions of the family, which include proper socialization, meeting fundamental life requirements, providing economic security, and reducing individual conflict. The belief that society is a system made up of a series of social subsystems that must be balanced and complimentary is central to proponents of functional constructionism [16].

The belief that society is a collection of regulated social systems, as well as the balance and complementarity of their components, is central to proponents of functional constructivism. The most essential aspect of this theory is that it analyzes family responsibilities and functions and emphasizes the family's valuable functions [17]. This indicates that any component of the family system whose functional role is broken, obstructed, or disrupted may eventually lead to dysfunction and family collapse. Therefore, early marriage may be a cause of family break-up if either spouse breaks up, fails to perform his or her functional role, or evades responsibility. According to the study [18], if this requirement is met, it is likely to result in increased psychological pressure on the spouses, or at least one of them, as well as a role

conflict, which can lead to acceptable practices such as multiple employment, or unethical or illegal practices to adapt to the new situation requiring each other's role, or to meet the personal or family needs that have been lost as a result. According to the researchers, this could lead to various forms of family disintegration, such as diversion and relocation of family members, particularly if the new means of adjustment require the obligation of caring for the family to be separated from its members.

2) *Conflict theory*: Conflict theorists argue that conflict is a continuous phenomenon that only ends with the demise of society and that conflict and the process of social change are thus inextricably linked [19]. But, since disagreements are inevitable in married life, may they be a source of family disintegration? According to the study [20], positive family differences can be seen as a means of obtaining rights and as a good way of improving the marital relationship, whereas negative conflicts lead to mismatch and dilution of marital relations, weakening the level of compatibility and harmony between spouses and eventually developing into situations of family disintegration. From this perspective, early marriage can be viewed as a cause of family disintegration if it results in negative conflicts, especially because there are many differences and conflicts associated with early marriages, often due to the selfishness and self-love of one or both of the immature spouses, the dysfunctional role of one of the spouses vis-à-vis the other or the children, or the difference in marital life reality from one or both of the immature spouses' expectations about rigor. According to numerous studies, the persistence of such negative conflicts has a negative impact on marital harmony and harmony situations and may eventually lead to family disintegration.

3) *Symbolic interaction theory*: According to this theory, the family plays an important role in preparing individuals for their future roles. The family's role is to instill in children a set of symbols, values, and standards. As a result, children can evaluate their actions and shape their future roles [7]. These symbols and meanings differ from family to family, and the individual first attempts to absorb the role expected of him before attempting to adapt a course based on his experiences and knowledge. The individual's daily interactions in his life, the circumstances surrounding his upbringing, and the foundation of symbols he learns from his family are among the most important factors influencing the construction of his personality [21]. According to this viewpoint, early marriage can be regarded as one of the causes of family disintegration in two cases: (1) if one or both spouses are unable to build a sufficient knowledge base that allows each of them to cope with married life and develop a common language with each other; or (2) if one or both spouses are unable to succeed in instilling sound educational values in their children, develop their feelings of love and loyalty to society, and develop their skills and abilities to interact with the members of society in a manner that enhances their integration and cohesion.

D. The Social Factors of Family Disintegration Related to Early Marriage

According to those points of view and based on studies on early marriage and previously analyzed theories, marriage requires a degree of love, desire, a sense of security, and the ability to properly raise and raise children, which can only be achieved in the case of so-called [19] cognitive readiness, which many boys and girls who marry early do not achieve. As a result, ignorance of marital life and adjustment requirements endangers marital peace and harmony (F1) and leads to family breakdown [20]. Furthermore, selfishness, selfishness, and the inability of some child couples to meet the requirements and duties of marriage and upbringing have caused family conflicts [22], which have sometimes escalated to the use of physical or verbal violence (F2), which is considered unacceptable in Yemeni society and is a major cause of divorce [8]. Authors in [9] goes on to say that the proper child rearing necessitates adequate knowledge of the foundations of sound socialization, health and psychology, and basic rules of care, and that one minor spouse's lack of such knowledge causes a dysfunctional role for the other or for the children [11] (F3), which can lead to family disintegration and disruption. On the other hand, that study [11] also, emphasized that the family was responsible for meeting the emotional needs of their sons and daughters, such as kindness, compassion, love, and justice, as well as releasing them from fear, anxiety, and anything else that might jeopardize their social and psychological well-being. According to the study, underage wives were found to be ineffective in this capacity [22]. Failure of minor spouses to fulfill their functional role in satisfying children's demand for family attachment (F4) is counterproductive, according to [19], resulting in emotions of alienation rather than feelings of family and community in children.

Locally, the study [23] indicates the existence of an expulsion relationship between early marriage and divorce, in which some minor spouses are subject to their guardians' desire to marry and choose the other partner, which causes a situation of incompatibility between the spouses, sometimes leading to separation (F5). Also, there have been a few instances when young children have been paired with adult females for various social reasons. Regardless of the reasons or motives, the age gap between them plays a crucial role in the incompatibility and harmony between the couples. In some of these circumstances, however, repeated marriages while keeping the first wife is a socially acceptable approach. Multiple marriages with the abandonment of the first wife (F6), on the other hand, is a bad outcome.

E. Health Factors of Family Disintegration and their Relationship to Early Marriage

Pregnancy and childbirth, in terms of health, are a substantial risk to the lives of teenage mothers [11] and can result in death, which is considered a cause of family breakdown. Some early marriages result in "early pregnancy," which leads to an increase in the number of underage mothers dying during childbirth (F7), has a negative impact on the mother's health, and even causes a variety of disorders (F8). In this context, the survey found [23] that two-thirds of Yemeni

moms had their babies aborted during their first trimester in 2008. According to the same survey, underage wives accounted for 19% of all moms who died during childbirth.

Psychologically, one of the minor spouses' incapacity to fulfill his functional role toward the other or to meet the needs of the children, as well as the consequent negative confrontations, leads to the escalation of psychological problems (F9) [20], [24].

III. METHODOLOGY

A. The Fuzzy Delphi Method

At the beginning of the sixties of the last century, the Delphi method was developed by two world scientists, "Olaf Helmer and Norman Dalkey" [25]. It is considered not only one of the most widely used and reliable surveying and expert judgment collection methods [26], [27], but one of the most widespread methods for solving numerous group decision-making problems by selecting and/or ranking factors, criteria, questionnaire elements, or measuring index elements [28]. Among the most prominent features of this method are that responses collected during its implementation remain uncharted and unknown, rely on a conditional phased statistical processing operation, and on countable, limited, and repeated processes that are managed and controlled by a phased outcome-based feedback operation. Besides, its outputs constitute consistent, revised, and collective statistical scores. Moreover, it is also characterized by the ability to address qualitative nature-based problems by relying on multiple survey rounds, which helps researchers in formulating additional quantitative survey rounds and promoting consensus opinions and effective decisions. So, it has been widely used to obtain a sequential series of consistent and revised responses and answers through multiple-round expert-opinion-based surveys in a lot of multidisciplinary studies [26], [27], [28].

TABLE I. THE PRIMARY FAMILY DISINTEGRATION FACTORS MOST ASSOCIATED WITH THE EARLY MARRIAGE PHENOMENON

Type	Factor	Description
Social factors (SF)	F1	Weakness in marital harmony and compatibility.
	F2	Increasing the domestic violence rates
	F3	The imbalance of the spouses' roles towards each other and towards their children.
	F4	Parents' failure in the proper family upbringing of their children.
	F5	Increasing divorce rates in marital cases that do not take place according to the common desire of the spouses.
	F6	Increasing of abandonment cases as a result of the heterogeneity arising from the presence of large age differences between spouses.
Health Factors (HF)	F7	Increasing the death rate among underage mothers during childbirth.
	F8	Negatively impacts on the health status of mothers
	F9	Increasing the rates of psychological cases among underage wives

While the efficiency of this technique and the widespread use of it over time have been proven, the inappropriateness of its application for many group decision-making situations has also been proven [29], [30]. For instance, some of these scenarios are: (1) limited time and financial resources of researchers due to the multiplicity of rounds imposed by the phased outcomes of surveys, especially when the assessment processes are conducted in fuzzy environments by a large number of heterogeneous experts, which requires more survey rounds to reach an acceptable convergence level in experts' prediction values, causing an increase in effort and costs. (2) cases that necessitate the use of uncommon and time-sensitive experts. In such multi-round survey situations, a shortage in the number of experts in the advanced survey rounds or a loss of the phased assessment data of some of them could occur, which also negatively affects the quality of the results. Furthermore, (3) evaluation cases that occur in ambiguous environments where conducting the evaluation process quantitatively is practically impossible due to a clear and unified understanding of measuring level. Rather, experts are forced to express their opinions through qualitative measures, which often have different meanings, connotations, and interpretations. Because this strategy is unable to deal with such uncertainty, it may produce erroneous results, lowering the quality of final decisions.

In any case, the current study is characterized by the above three cases, which means that this technique in its traditional version is not suitable for use. In this regard, the fuzzy version of this method has been developed by [31], and its calculation procedures rely on fuzzy numbers and allow experts to express their opinions through them. It is also able to deal with the uncertainty problem on the one hand. On the other hand, it relies on a limited number of survey rounds, which contributes to reducing the costs, effort, and time of researchers and experts at the same time. It also enhances an expert's interest and desire for continuity, which contributes to raising the recovery rates of their questionnaires. It helps to enhance the completeness and consistency of opinions as it provides a mechanism to deal with non-consensus cases and obtain consensus in the opinions without subjecting the original opinions of experts to change, which also leads to more realistic and objective decisions.

B. FDM Implementation Procedures

Given that decision-making processes require a good study and analysis of decision requirements [32], [33], [34],[35] and the selection and application of the appropriate systematic tools to solve them [36], [37], [38],[39] and based on the recommendations of previous studies that have demonstrated the effectiveness of Fuzzy based techniques in general [40],[41] compared to traditional evaluation techniques [42],[43] and taking into account the efficiency of the Fuzzy version of the Delphi method, this study relied on this method to solve the study's problem and on the following implementation steps [30] [44] [45]:

Step 1: Determining the main areas of the factors to be evaluated according to the nature of the study.

Step 2: Reviewing pertinent theories and literature and suggesting significant factors for each main area.

Step 3: Developing the data collection tool, selecting the panel of experts, and gathering judgments of the decision group on each factor: Determine the assessment score of each factor's importance as presented by each of them through the application of a five-point Likert variable and convert the collected scores to their equivalent fuzzy numbers.

Step 4: Data Processing and Analysis:

Step 4-1: Calculating the average fuzzy rating scores of group decisions for each alternative factor.

As noted above, the evaluation process of the causes of family disintegration is usually accompanied by a case of ambiguity, which requires a conversion of the assessment data obtained using Likert's five-point evaluation scale into equivalent fuzzy numbers within a specific fuzzy logic set. In this regard, the literature provides many forms of those numbers, such as triangular and trapezoidal numbers. The fuzzy set in the subset of real numbers (X) is defined by a two-part combination, the first representing the "x" component, while the second reflects the degree to which that element belongs to the fuzzy set. A numerical membership function (MF) is used to determine whether an element belongs or does not belong to that fuzzy set, and its values are limited to the range [0,1]. The closer this value is to zero, the less the fact that the element belongs to the set, and the opposite is true, as the validity of that statement increases the closer the value of the function is to one.

Given that the personal opinions of experts about the appropriateness of a particular factor to be an indicator of a general domestic framework for measuring, ranking, and structured comparing of family disintegration causes associated with early marriage in the Republic of Yemen are relative and probabilistic and cannot be determined accurately and objectively, the use of MF functions constitutes an appropriate and acceptable solution to deal with this case of ambiguity. As shown in Table II, fuzzy numbers with relative values limited to the range [0,1] were used in this study to represent cases of ambiguity among respondents.

However, the analysis process requires calculating the average fuzzy rating scores of group decisions for each alternative factor. Assuming the assessment value of the appropriateness of No. "z" factor given by an expert No. "r" of a total number "n" experts is $\tilde{w}_{rz} = (a_{rz}, b_{rz}, c_{rz})$, $r = (1.2 \dots n)$, $z = (1.2 \dots m)$.

Then the average fuzzy number \tilde{w}_z of No. "z" factor is defined as [44], [45]:

$$\tilde{w}_z = (a_z, b_z, c_z) = \left(\frac{1}{n} \sum_{r=1}^n a_{rz}, \frac{1}{n} \sum_{r=1}^n b_{rz}, \frac{1}{n} \sum_{r=1}^n c_{rz} \right) \quad (1)$$

TABLE II. LIKERT AND FUZZY SCORING SCALES

Likert scale scoring	Linguistic variable	Fuzzy Scale scoring
1	Highly Not Agree	(0,0,0,0,2)
2	Not Agree	(0,0,0,2,0,4)
3	Moderately / Not sure	(0,2,0,4,0,6)
4	Agree	(0,4,0,6,0,8)
5	Highly Agree	(0,6,0,8,1,0)

Step 4-2: Defuzzification: convert the calculated average fuzzy rating score \widetilde{w}_z of each factor "z" to its equivalent crisp numbers.

In this study, the simple center of gravity approach was applied to defuzzify the aggregated fuzzy rating scores (D_z) of factors as follows.

$$D_z = \frac{(a_z \cdot b_z \cdot c_z)}{n} \quad (2)$$

Step 5: Examine the acceptability of the evaluation domain.

Step 5-1: For each factor "z", calculate the difference value (D_{rz}) between the average fuzzy number (\widetilde{w}_z), and each expert's fuzzy evaluation value (\widetilde{w}_{rz}) using equation 3.

$$D_{rz} = \sqrt{\frac{(a_z - a_{rz})^2 + (b_z - b_{rz})^2 + (c_z - c_{rz})^2}{3}} \quad (3)$$

Step 5-2: Determining the threshold value (Th_z) of each factor "z" by using equation 4.

$$Th_z = \frac{1}{n} \sum_{r=1}^n D_{rz} \quad (4)$$

Step 5-3: Testing the threshold value (Th_{domain}) of each evaluation domain by using equation 5.

$$Th_{domain} = \frac{1}{n} \sum_{z=1}^m Th_z \quad (5)$$

Based on the " Th_{domain} " value, the acceptability of the evaluation domain should be determined. In this study, an evaluation domain is accepted if $Th_{domain} \leq 0.02$.

Step 6: Testing the Expert Group Consensus:

Using equation 6, calculate the expert agreement on each evaluated factor.

$$EA_z = \frac{E_z}{n} \% \quad (6)$$

Where the " E_z " is the total number of experts, who's the distance between their fuzzy evaluation values (\widetilde{w}_{rz}) on a particular factor "z" and the average fuzzy number of all experts on that factor (\widetilde{w}_z) is ≤ 2 . Based on an expert's agreement value " EA_z ", the expert group consensus for each factor should be determined. In this study, an expert's agreement with a " $(EA_z) \geq 75\%$ " is used to screen out the expert group consensus for each factor. Factors with an expert consensus of less than 75% are ignored.

Step 6: Testing the final acceptability decision of each evaluation factor.

Once the evaluation domain is accepted and the list of factors with an acceptable expert consensus level is defined. The selected accepted factors should be analyzed based on the overall group evaluation crisp scores (D_z). These factors should be ranked, and the low-rated (low-ranked) factors should be discarding. In this study, the value of "0.4" was applied to describe the low-ranked factors.

C. Assessment Model of Study

Based on the previously described implantation stages of the Fuzzy Delphi Method, the assessment model for

developing the general framework's indicators to evaluate, analyze, and compare the family disintegration causes associated with the early marriage phenomenon in the Republic of Yemen was designed and applied. Fig. 1 describes it.

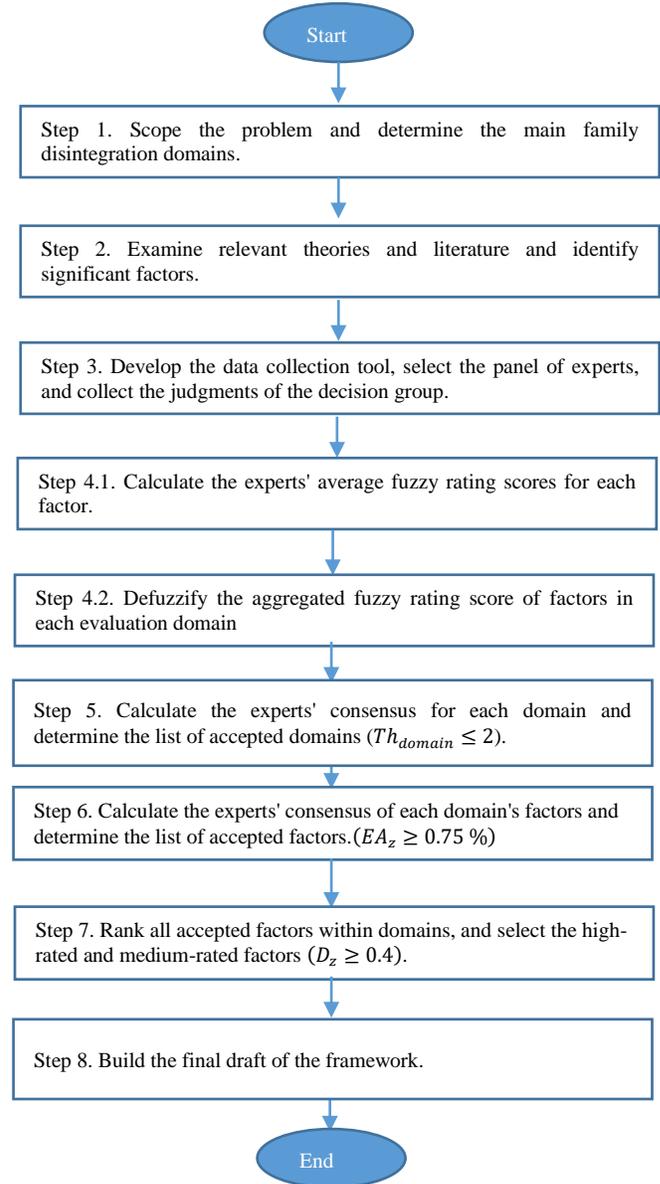


Fig. 1. The Assessment Model of Study.

IV. ASSESSMENT MODEL APPLICATION AND RESULTS

A. Scoping the Problem

The social and health domains of the family disintegration factors that are related to the phenomenon of family disintegration were chosen as the two main areas of factors to be identified and prioritized (refer to Section 1).

B. Reviewing Pertinent Theories and Literature and Suggest the Significant Factors

Nine major factors of family disintegration that are highly related to early marriage have been proposed, six of which are

social factors, and three others are health. These factors were proposed based on reviewing more than thirty relevant studies (refer to Section 2), accordingly, the primary family disintegration factors most associated with the early marriage phenomenon were crystallized, as shown in Table I.

C. Developing the Data Collection Tool, Selecting the Panel of Experts and Collecting Judgments of the Decision Group

This study followed the recommendations of previous studies on the application of FDM, whereby at least 10 experts must be selected in a purposeful manner that meets the requirements of homogeneity in order to participate in the evaluation [30], [44], [45]. To achieve the requirements of homogeneity, this study followed the recommendations of the study [30], and a total of twenty-four experts have been selected, all of whom are currently engaged in thematic work close to the topic of this study and have accumulated knowledge and experience for more than 10 years in a related domain. Depending on the specialization, they are equally split up into three directions: demography science, sociology, and health and psychological sciences. This study used an assessment tool (questionnaire) to collect data from experts. A number of 24 closed questionnaires were sent by e-mail to all experts, each questionnaire containing nine factors, and they were asked to indicate their level of agreement on the

appropriateness of these factors as indicators of a general domestic framework for measuring, ranking, and structured comparing of family disintegration associated with early marriage in the Republic of Yemen, using the five-point Likert scale (strongly disagree = (1), strongly agree = (5)). Subsequently, all questionnaires were also received by e-mail. After that, all ratings were converted to their equivalent fuzzy number scores using the conversion table proposed by [44] (see Table II). Table III illustrates the fuzzy ratings of all experts.

D. Data Processing and Analysis

For this purpose, a simulation program has been developed using the Excel 2013 application. Using that tool, the average fuzzy rating score of all experts (\widetilde{w}_{rz}) on each factor (z) was calculated (\widetilde{w}_z) and defuzzified (D_z) using equations (1) and (2), respectively.

These two rating score vectors are illustrated in Table IV. Subsequently, the experts' consensus of the two evaluation domains of study (Th_{domain}) was examined using equations (3), (4), and (5). Both domains were accepted ($Th_{domain} \leq 0.02$). After that, the experts' consensus on each evaluation factor was tested using equation (6). This step adopts factors with a threshold (EA_z) above or equal to 75%

TABLE III. EXPERTS' FUZZY RATING SCORES

Expert	F1	F2	F3	F4	F5	F6	F7	F8	F9
1	(0.4,0.6,0.8)	(0.4,0.6,0.8)	(0.4,0.6,0.8)	(0.4,0.6,0.8)	(0.6,0.8,1.0)	(0.0,2,0.4)	(0.4,0.6,0.8)	(0.6,0.8,1.0)	(0.4,0.6,0.8)
2	(0.4,0.6,0.8)	(0.4,0.6,0.8)	(0.4,0.6,0.8)	(0.4,0.6,0.8)	(0.6,0.8,1.0)	(0.2,0.4,0.6)	(0.4,0.6,0.8)	(0.4,0.6,0.8)	(0.4,0.6,0.8)
3	(0.4,0.6,0.8)	(0.4,0.6,0.8)	(0.4,0.6,0.8)	(0.4,0.6,0.8)	(0.6,0.8,1.0)	(0.0,0.2,0.4)	(0.4,0.6,0.8)	(0.4,0.6,0.8)	(0.4,0.6,0.8)
4	(0.4,0.6,0.8)	(0.0,0.2,0.4)	(0.4,0.6,0.8)	(0.4,0.6,0.8)	(0.6,0.8,1.0)	(0.0,0.2,0.4)	(0.4,0.6,0.8)	(0.6,0.8,1)	(0.6,0.8,1.0)
5	(0.4,0.6,0.8)	(0.4,0.6,0.8)	(0.4,0.6,0.8)	(0.4,0.6,0.8)	(0.6,0.8,1.0)	(0.0,0.2,0.4)	(0.4,0.6,0.8)	(0.4,0.6,0.8)	(0.4,0.6,0.8)
6	(0.4,0.6,0.8)	(0.4,0.6,0.8)	(0.4,0.6,0.8)	(0.4,0.6,0.8)	(0.6,0.8,1.0)	(0.2,0.4,0.6)	(0.4,0.6,0.8)	(0.4,0.6,0.8)	(0.4,0.6,0.8)
7	(0.4,0.6,0.8)	(0.0,6,0.8,1)	(0.4,0.6,0.8)	(0.4,0.6,0.8)	(0.6,0.8,1.0)	(0.2,0.4,0.6)	(0.6,0.8,1.0)	(0.6,0.8,1)	(0.4,0.6,0.8)
8	(0.4,0.6,0.8)	(0.4,0.6,0.8)	(0.4,0.6,0.8)	(0.4,0.6,0.8)	(0.6,0.8,1.0)	(0.0,0.2,0.4)	(0.4,0.6,0.8)	(0.4,0.6,0.8)	(0.4,0.6,0.8)
9	(0.4,0.6,0.8)	(0.4,0.6,0.8)	(0.4,0.6,0.8)	(0.4,0.6,0.8)	(0.4,0.6,0.8)	(0.2,0.4,0.6)	(0.4,0.6,0.8)	(0.4,0.6,0.8)	(0.4,0.6,0.8)
10	(0.4,0.6,0.8)	(0.6, 0.0,8,1)	(0.4,0.6,0.8)	(0.4,0.6,0.8)	(0.4,0.6,0.8)	(0.2,0.4,0.6)	(0.4,0.6,0.8)	(0.4,0.6,0.8)	(0.4,0.6,0.8)
11	(0.4,0.6,0.8)	(0.6,0.8,1.0)	(0.4,0.6,0.8)	(0.4,0.6,0.8)	(0.4,0.6,0.8)	(0.0,0.2,0.4)	(0.6,0.8,1.0)	(0.6,0.8,1.0)	(0.6,0.8,1.0)
12	(0.4,0.6,0.8)	(0.4,0.6,0.8)	(0.4,0.6,0.8)	(0.6,0.8,1.0)	(0.4,0.6,0.8)	(0.2,0.4,0.6)	(0.6,0.8,1.0)	(0.6,0.8,1.0)	(0.4,0.6,0.8)
13	(0.4,0.6,0.8)	(0.6,0.8,1.0)	(0.4,0.6,0.8)	(0.2,0.4,0.6)	(0.6,0.8,1.0)	(0.2,0.4,0.6)	(0.4,0.6,0.8)	(0.4,0.6,0.8)	(0.4,0.6,0.8)
14	(0.0,0.2,0.4)	(0.4,0.6,0.8)	(0.4,0.6,0.8)	(0.4,0.6,0.8)	(0.4,0.6,0.8)	(0.0,0.2,0.4)	(0.4,0.6,0.8)	(0.4,0.6,0.8)	(0.4,0.6,0.8)
15	(0.4,0.6,0.8)	(0.0,0.2,0.4)	(0.4,0.6,0.8)	(0.4,0.6,0.8)	(0.6,0.8,1.0)	(0.2,0.4,0.6)	(0.6,0.8,1.0)	(0.6,0.8,1.0)	(0.4,0.6,0.8)
16	(0.4,0.6,0.8)	(0.4,0.6,0.8)	(0.4,0.6,0.8)	(0.2,0.4,0.6)	(0.4,0.6,0.8)	(0.2,0.4,0.6)	(0.4,0.6,0.8)	(0.4,0.6,0.8)	(0.4,0.6,0.8)
17	(0.4,0.6,0.8)	(0.4,0.6,0.8)	(0.4,0.6,0.8)	(0.2,0.4,0.6)	(0.4,0.6,0.8)	(0.2,0.4,0.6)	(0.6,0.8,1.0)	(0.6,0.8,1.0)	(0.6,0.8,1.0)
18	(0.4,0.6,0.8)	(0.2,0.4,0.6)	(0.2,0.4,0.6)	(0.6,0.8,1)	(0.6,0.8,1.0)	(0.2,0.4,0.6)	(0.6,0.8,1.0)	(0.4,0.6,0.8)	(0.4,0.6,0.8)
19	(0.4,0.6,0.8)	(0.4,0.6,0.8)	(0.4,0.6,0.8)	(0.4,0.6,0.8)	(0.6,0.8,1.0)	(0.0,0.2,0.4)	(0.6,0.8,1.0)	(0.6,0.8,1.0)	(0.6,0.8,1.0)
20	(0.4,0.6,0.8)	(0.2,0.4,0.6)	(0.0,0.0,0.2)	(0.4,0.6,0.8)	(0.6,0.8,1.0)	(0.2,0.4,0.6)	(0.2,0.4,0.6)	(0.2,0.4,0.6)	(0.4,0.6,0.8)
21	(0.4,0.6,0.8)	(0.4,0.6,0.8)	(0.2,0.4,0.6)	(0.4,0.6,0.8)	(0.2,0.4,0.6)	(0.0,0.2,0.4)	(0.6,0.8,1.0)	(0.4,0.6,0.8)	(0.6,0.8,1.0)
22	(0.4,0.6,0.8)	(0.4,0.6,0.8)	(0.4,0.6,0.8)	(0.2,0.4,0.6)	(0.6,0.8,1.0)	(0.2,0.4,0.6)	(0.6,0.8,1.0)	(0.6,0.8,1.0)	(0.4,0.6,0.8)
23	(0.4,0.6,0.8)	(0.2,0.4,0.6)	(0.4,0.6,0.8)	(0.4,0.6,0.8)	(0.6,0.8,1.0)	(0.4,0.6,0.8)	(0.2,0.4,0.6)	(0.2,0.4,0.6)	(0.6,0.8,1.0)
24	(0.4,0.6,0.8)	(0.4,0.6,0.8)	(0.4,0.6,0.8)	(0.2,0.4,0.6)	(0.6,0.8,1.0)	(0.0,0.2,0.4)	(0.4,0.6,0.8)	(0.4,0.6,0.8)	(0.4,0.6,0.8)

TABLE IV. THE FUZZY AVERAGE EVALUATION SCORES (\bar{w}_z), AND THEIR EQUIVALENT DEFUZZIFIED (D_z) VALUES

Type	F	\bar{w}_z	D_z
SF	F1	(0.383,0.583,0.783)	0.58
	F2	(0.375,0.575,0.775)	0.57
	F3	(0.367,0.558,0.758)	0.56
	F4	(0.375,0.575,0.775)	0.57
	F5	(0.525,0.725,0.925)	0.72
	F6	(0.125,0.325,0.525)	0.32
HF	F7	(0.458,0.658,0.858)	0.58
	F8	(0.458,0.658,0.858)	0.57
	F9	(0.450,0.650,0.850)	0.56

As a result, all factors have been accepted at this stage. Then, factors were ranked in accordance with the defuzzified assessment values (D_z), and only the high-rated and medium-rated factors were selected.

The ranking step uses factors with a defuzzified value of 0.4 or above, and those factors with a defuzzified value of less than 0.4 are ignored. However, with an average group assessment rate of more than 0.4 and an average group evaluation consistency rate of more than 75%, eight out of all nine evaluated factors were agreed to be general local framework's indicators for measuring, ranking, and structured comparing of family disintegration associated with an early marriage phenomenon in the Republic of Yemen, while, with the an acceptable consistency rate ($EA_z=96\%$) and an average group assessment rate of ($D_z=0.325<=0.4$), the last social factor (the sixth factor) was agreed to be inappropriate for that.

Table V shows the obtained consistency ratios of the evaluation domains (Th_{domain}) and factors (EA_z), the defuzzified evaluation scores (D_z) and levels (EL) of each factor, the accepted and deleted factors (Status), and the ranking order of accepted factors (R).

TABLE V. EVALUATION RESULTS

Type	Th_{domain}	F	D_z	EL	EA_z	Status	R
SF	0.081	F1	0.58	Moderate	96%	ACCEPTED	4
		F2	0.57	Moderate	75%	ACCEPTED	5
		F3	0.56	Moderate	96%	ACCEPTED	6
		F4	0.57	Moderate	92%	ACCEPTED	5
		F5	0.72	high	96%	ACCEPTED	1
		F6	0.32	Low	96%	REJECTED	
HF	0.096	F7	0.66	high	92%	ACCEPTED	2
		F8	0.66	high	92%	ACCEPTED	2
		F9	0.65	high	100%	ACCEPTED	3

V. DISCUSSION OF RESULTS

As mentioned earlier, Yemen's political conditions and the associated social, health, and economic changes have caused

domestic instability in Yemen. However, putting in place the necessary measures, solutions, and treatments, and distributing resources and services that contribute to limiting their effects, requires carrying out many relevant studies and analyses. Some examples of these requirements are: evaluation processes for the causes of disintegration; classification, clustering, and ranking of causes in each sector according to the probability and impact of each of the causes; classification and clustering of geographical sectors according to the factor's influence level on them and according to their need for relevant solutions and treatments; and determining the percentage of what each sector needs compared to others. Weaknesses in one or several areas related to the factors of disintegration.

In any case, these processes, which we briefly refer to in this study as assessment, arrangement, classification, and regular comparison, can only bear fruit by building the correct assessment framework, which includes comprehensive and accurate measurement indicators. So, this study aimed to build a general local framework that can later be used as a main tool for the implementation of these processes in order to solve this problem.

However, this paper fulfilled the purpose of the study. And the general local framework's indicators for measuring, ranking, and structured comparing of family disintegration associated with the early marriage phenomenon in Yemen was built by getting the consensus of a group of local experts on the relevance of the theoretically predefined factors using the Fuzzy Delphi Method.

The findings disclosed in the current paper indicate the consistency of experts' opinions on all indicators. The consistency rates across those indicators are acceptable (greater than 75%). A small degree of discrepancy (not exceeding 3%) between the experts' consistency ratio averages on the health (95%) and social (92%) fields' indicators was observed. This confirms the accuracy and integrity of the theoretical procedures followed for theoretically extracting the initial indicators of study. It also confirms the accuracy of the methodological procedures used in selecting a homogeneous sample of experts [41], and the accuracy and objectivity of experts during the evaluation process.

However, the expert opinions' consistency on indicators does not mean that all of them are appropriate. For example, with an average evaluation rate of 0.33 and a very high consistency rate of 96%, experts unanimously agreed on the poor suitability of the sixth indicator, "Incrementing of abandonment cases as a result of heterogeneity arising from the presence of large age differences between spouses," which represents 11% of all initial indicators, to represent a measurement indicator of family disintegration resulting from the phenomenon of early marriage. And this can be explained by the fact that this indicator is more related to the phenomenon of polygamy than to the phenomenon of early marriage. According to [46], this case is related to wives who marry boys younger than them and with large age differences, where large age differences over time cause cases of polygamy, especially in light of the unequal relationship between them.

In addition, for various reasons, which may be economic or social, some of these cases may result in the husband preferring to live with the second wife and abandoning the first wife, which may sometimes lead to divorce [47]; So, the sixth social factor may be considered a cause of family disintegration, but it is not mainly related to early marriage but rather to the phenomenon of multiple marriage.

Regarding the accepted indicators, with the exception of the sixth social indicator, all preliminary social and health indicators were accepted.

With an evaluation average of 0.72, the fifth indicator, "Increasing divorce rates in marital cases that do not take place according to the common desire of the spouses," came in first place. And this may be due to the local prevalence rate of these marriage cases [18], on the one hand, and their negative social effects on the compatibility and homogeneity between spouses [48], and their causing dire social consequences that reduce the cohesion of local families [49], on the other hand.

On the health side, with an average group evaluation score of 0.66, the seventh and eighth indicators, "increase in the number of deaths of underage mothers during childbirth" and "impact on maternal health" came in second place, and with a very small difference (-0.008), the ninth indicator, "Increasing the rates of psychological cases among underage wives," came in third place, which indicates the importance of the health factors of early marriage and the extent to which they cause family disintegration. And this may be due to the negative health effects of child marriage [6] and the resulting dire social consequences that hinder the wife from playing her integrative role in society [23].

These findings also emphasize the magnitude of health problems associated with early marriage, such as mental illness, personality disorders, disorders in sexual relations, depression and anxiety, increasing rates of induced abortions, and increasing rates of childbirth [4]. According to the International Women's Health Coalition (IWHC), women married before the age of 15 are "five times more likely to die in childbirth than women in their 20s and face a higher risk of pregnancy-related injuries [4]. Also, these findings highlight the need to consider health factors during the relevant monitoring, evaluation, and planning processes.

Socially, the average evaluation degree of the accepted social indicators (0.603) was slightly and almost insignificantly lower than its equivalent for health indicators (0.656), which also highlights the importance of indicators of a social nature within the proposed framework. In addition, with relatively different evaluation averages swinging between 0.58 and 0.57, the first four social indicators were ranked in the fourth, fifth, sixth, and fifth positions, respectively. And this confirms that there are no significant differences in their importance among experts.

Based on the foregoing, with the exception of the sixth factor, the experts' assessments confirmed that all factors that were theoretically derived could be considered as appropriate indicators to measure the family disintegration associated with the child marriage phenomenon in Yemen, and that they could be practically used to measure the causes of disintegration in

different regions of Yemen. Also, previous local and global studies such as [40, 50,51] on the development of planning and resource allocation decision support systems have confirmed that evaluation frameworks are a key component of the evaluation, ranking, comparison, classification, and clustering models of the national planning and resource allocation decision support systems, which are used to determine the actual needs of the geographical sectors; arrange and classify sectors according to their needs; and arrange and classify the causes according to the level of their spread and impact and the percentage of needs needed for each sector in comparison with other sectors.

These studies also indicate that these systems effectively contribute to the optimal utilization of resources, the promotion of sustainability practices, and the requirements of justice and the equal allocation of resources to address them. Taking into account all these assumptions, the researchers in this study recommend that decision makers adopt the proposed framework as part of the national decision support system for family and social stability planning and resource allocation management.

VI. LIMITATIONS, APPLICATIONS AND FUTURE WORK

The study addressed the development of indicators within the general framework that could be applied in practice for other research purposes and examined only social and health indicators of family disintegration linked specifically to the phenomenon of early marriage in Yemen and ignored other types of indicators or factors, such as social and health indicators associated with polygamy or economic factors associated with family disintegration in general. Also, the indicators' ranking process has been carried out from the perspective of their suitability to measure the extent to which early marriage in Yemen generally causes family disintegration. In other words, priorities were not studied from the perspective of their impact level on the family disintegration state, nor from their relative importance as influential criteria according to certain developmental considerations.

Nevertheless, the proposed framework is fundamental and central to resolving many of the resolution's issues that this study has not been able to address. For example, to prioritize the proposed framework's domains and indicators as benchmarks from different experts' specialized perspectives and according to specific developmental considerations through the use of appropriate techniques such as AHP, F-AHP, BWM, or F-BWM; Conduct an analysis of the causal analysis between indicators, and to understand the different levels of impact results and the logical relationship that links them to each other, with the possibility of studying them from a general perspective, or from multiple specialized perspectives for decision experts through the application of appropriate techniques such as ANP, F-ANP, DEMATEL, or F-DEMATEL.

On the other hand, it could be used to assess and rank the causes of family disintegration in different Yemeni regions from the perspective of members of the community in those areas using a suitable hybrid technique that combines a subjective weighting method or/and objective weighting method with another ranking method such as TOPSIS or F-

TOPSIS; or to carry out structural comparative studies between those indicators in the different Yemeni governorates or areas to propose more effective solutions and treatments. This will help to promote the equity and investment of government resources and budgets on the one hand, and to attract institutions with different directions to implement their social responsibility practices in line with their directions on the other. Based on the foregoing, these applications will constitute the most prominent set of future research directions for researchers in this study.

VII. CONCLUSION

The problem of family disintegration associated with the phenomenon of early marriage is one of the most prominent local issues negatively affecting the stability, prosperity, and development of society. The planning and building of sustainable sound strategies to reduce the negative effects of this problem requires the building of integrated decision support systems, capable of assisting in the implementation of assessments, arrangement, classification, comparison, and other relevant planning functions commensurate with local conditions, environment, and priorities. Also, the effectiveness, quality, and functionality of such systems depend on the validity and efficiency of the evaluation tools used to collect the inputs of such systems, which in turn must include real indicators that have been identified in accordance with the relevant theoretical foundations and principles and through the real participation of experts and specialists, taking into account the requirements and variables imposed by the local environment.

To achieve that goal, the problem was first examined and described as a decision problem. This step was followed by the selection and analysis of a set of relevant literature and theories. At this stage, a total of six social and three health indicators were defined to form the initial social and health indicators of the study. Then, twenty-four local experts were surveyed through a questionnaire tool designed for this purpose. After that, the FDM methodology was applied to study and analyze experts' opinions on the appropriateness of developed indicators. Based on that, eight key final indicators of family disintegration were drawn to represent the main structure of the proposed framework.

By analyzing the study's findings, the following conclusions were reached: (1) Experts focused on the subject of divorce. They believe that divorce arising from cases where the spouses or one of them is forced to marry another without the full conviction or desire to associate with them, to satisfy the family or to fulfill their wishes, is the main indicator for measuring family disintegration in early marriages. (2) Experts focused heavily on the health criterion, with an average assessment of the indicators of this criterion of 0.656, possibly because their concerns were focused on the importance of women's safety and health, which is a key pillar of family stability. (3) Although the average evaluation of health domain indicators was higher than social, the difference was small and did not exceed 0.05, which also underscores the importance of this domain. (4) Experts pay almost equal attention to health indicators, and relatively different to social indicators. The value of the average standard deviation of the

health indicators assessment averages was (0.0028), and that value increased by almost twenty-one times for the social standard indicators (0.059). Despite that, the discrepancy in the importance of social indicators for the group of experts remains relative and ineffective.

In any case, this study reached one final output called the general framework for the evaluation of the family disintegration causes associated with the child marriage phenomenon in Yemen. This framework is characterized by specialization, systematic, and appropriateness, as it was built according to multiple social and health theories, with the participation of local experts who live, influence, and are affected by the surrounding environmental conditions and using a reliable and globally approved decision support technique in order to contribute to solving a specific problem, according to the Yemeni context. Also, it has many real applications, and it can be used as a major component of the assessment, ranking, comparison, classification, and clustering models of the national and social related planning and resource allocation decision support systems.

ACKNOWLEDGMENT

This research was supported by the Deanship of Scientific Research at Prince Sattam Bin Abdulaziz University.

REFERENCES

- [1] G. G. Fierro and M. A. H. Martínez, "Leadership Competences Education for Social Development: a TEC21 Model in the i Semester Experience," *South Florida Journal of Development*, vol. 2, no. 2, pp. 1837–1852, May 2021.
- [2] T. Jarana-Díaz, M. Romero-Martín, J. A. Ponce-Blandón, and N. Jiménez-Picón, "Integrative Review of Related Factors and Defining Characteristics of Lack of Family Integrity," *International Journal of Nursing Knowledge*, vol. 32, no. 1, pp. 44–52, Jun. 2020.
- [3] L. Holtzhausen and E. Campbell, "Adverse Childhood Experiences as a Risk Factor for Anti-Social Behaviour Among Young Adults in the Western Cape, South Africa," *Acta Criminologica: African Journal of Criminology & Victimology*, vol. 34, no. 1, pp. 24–47, 2021.
- [4] A. Richter, "Forgotten Daughters: Child Marriage in Yemen Civil War," *The Florida State University*, 2022.
- [5] P. Villar, "Paternal mortality, early marriages, and marital trajectories in Senegal," *World Development*, vol. 142, p. 105421, Jun. 2021.
- [6] M. Lestari and S. E. Adiyatma, "Marriage Cancelled, What about the Rights for Children?," *The Indonesian Journal of International Clinical Legal Education*, vol. 2, no. 2, pp. 167–180, Jun. 2020.
- [7] K. Hunersen, B. Attal, A. Jeffery, J. Metzler, T. Alkibsi, S. Elnakib, and W. C. Robinson, "Child marriage in Yemen: a mixed methods study in ongoing conflict and displacement," *Journal of Refugee Studies*, Feb. 2021.
- [8] A. P. Thompson, E. Wine, S. E. MacDonald, A. Campbell, and S. D. Scott, "Parents' Experiences and Information Needs While Caring for a Child With Functional Constipation: A Systematic Review," *Clinical Pediatrics*, vol. 60, no. 3, pp. 154–169, Oct. 2020.
- [9] E. A. Sarfo, J. Salifu Yendork, and A. V. Naidoo, "Understanding Child Marriage in Ghana: The Constructions of Gender and Sexuality and Implications for Married Girls," *Child Care in Practice*, vol. 28, no. 2, pp. 228–241, Jan. 2020.
- [10] U. Zartler, "Children and parents after separation," *Research Handbook on the Sociology of the Family*, pp. 300–313, 2021.
- [11] P. Shahi, P. D. Tamang, P. Simkhada, and K. S. Rawat, "Child Marriage-Knowledge, practice and its attributed consequences among early married women in Jumla, Nepal," *Asian Pacific Journal of Health Sciences*, vol. 6, no. 1, pp. 140–148, Mar. 2019.
- [12] D. Davidson, "Sibling loss - disenfranchised grief and forgotten mourners," *Bereavement Care*, vol. 37, no. 3, pp. 124–130, Sep. 2018.

- [13] A. A. Al Shirawi and J. I. Alumran, "Work-Family Balance and its Relationship to Marital Adjustment among a Sample of Saudi Female Teachers," *Journal of Educational & Psychological Sciences*, vol. 20, no. 01, pp. 11–40, Mar. 2019.
- [14] S. M. A., Hamid, and K. H. Sachit, family Disintegration An analytical study. *Mustansiriyah Journal of Arts*, 43(87), pp.175-191, 2019.
- [15] A. Goodman, M. Snyder, and K. Wilson, "Exploring Indigenous youth perspectives of mobility and social relationships: A Photovoice approach," *The Canadian Geographer / Le Géographe canadien*, vol. 62, no. 3, pp. 314–325, Apr. 2018.
- [16] P. Sohlberg, "Functionalist Construction Work in Social Science," Mar. 2021.
- [17] M. D. L. F. Vargas, "The social sciences and humanities on family and kinship: contributions to the Family Health Strategy," *História, Ciências, Saúde-Manguinhos*, issue. 28, pp. 351-374, 2021.
- [18] Nawal Hamza, Suad Al-Habsi, Laila Al-Hazoura, Hanan Hajeb, "Early Marriage: A Study in the Concept, Causes and Effects, Higher Diploma Thesis, Sana'a University, 2008.
- [19] L. McDougal, E. C. Jackson, K. A. McClendon, Y. Belayneh, A. Sinha, and A. Raj, "Beyond the statistic: exploring the process of early marriage decision-making using qualitative findings from Ethiopia and India," *BMC Women's Health*, vol. 18, no. 1, Aug. 2018.
- [20] N. A. John, J. Edmeades, and L. Murithi, "Child marriage and psychological well-being in Niger and Ethiopia," *BMC Public Health*, vol. 19, no. 1, Aug. 2019.
- [21] S. Stryker, "Symbolic Interactionism: Themes and Variations," *Social Psychology*, pp. 3–29, Sep. 2017.
- [22] G. DeVos and H. Wagatsuma, "5. Family Life and Delinquency: Some Perspectives from Japanese Research," *Transcultural Research in Mental Health*, pp. 59–87, Dec. 2021.
- [23] Nawal Hamza, Suad Al-Habsi, Laila Al-Hazoura, Hanan Hajeb, "Early Marriage: A Study in the Concept, Causes and Effects, Higher Diploma Thesis, Sana'a University, 2018.
- [24] M. Firooze, "The role of psychologist in curbing psychological abnormalities in children of divorce," *Фундаментальные и прикладные исследования в современном мире*, 3(29), pp 52-56, 2021.
- [25] N. Dalkey and O. Helmer, "An Experimental Application of the DELPHI Method to the Use of Experts," *Management Science*, vol. 9, no. 3, pp. 458–467, Apr. 1963.
- [26] G. J. Skulmoski, F. T. Hartman, and J. Krahn, "The Delphi Method for Graduate Research," *Journal of Information Technology Education: Research*, vol. 6, pp. 001–021, 2007.
- [27] R. Boulkedid, H. Abdoul, M. Loustau, O. Sibony, and C. Alberti, "Using and Reporting the Delphi Method for Selecting Healthcare Quality Indicators: A Systematic Review," *PLoS ONE*, vol. 6, no. 6, p. e20476, Jun. 2011.
- [28] B. D. Lund, "Review of the Delphi method in library and information science research," *Journal of Documentation*, vol. 76, no. 4, pp. 929–960, Feb. 2020.
- [29] Adel A. Nasser, Mohammed M Said, Mijahed N. Aljober "Application of Selected MCDM Methods for Developing a Multi-Functional Framework for Eco-Hotel Planning in Yemen," *International Journal of Computer Sciences and Engineering*, vol. 9, no. 10, pp. 7–18, sept. 2021.
- [30] Mohammed M Said, Adel A Nasser and Abdualmajed A Alkhalaidi. , "Prioritization of the Eco-hotels Performance Criteria in Yemen using Fuzzy Delphi Method," *International Journal of Applied Information Systems* 12(36):20-29, March 2021.
- [31] A.Ishikawa, M. Amagasa, T. Shiga, G. Tomizawa, R.Tatsuta, and H. Mieno, , "The max–min Delphi method and fuzzy Delphi method via fuzzy integration," *Fuzzy Sets and Systems*, vol 55 , pp. 241–253, 1993.
- [32] А.А.Гуламов, С.Н. Михайлов, А. А. Насер, " Модель Процессов Информационно-Аналитического Обеспечения Научных Исследований Вуза ," *Информационно-Измерительные И Управляющие Системы*. Vol. 9. no. 4, pp.. 28-31, 2011.
- [33] А. А. Насер, А. А. Гуламов , "Информационная Модель Процессов Информационно-Аналитического Обеспечения В Вузе, Вести Высших Учебных Заведений Черноземья. vol 2 (24). pp. 111-114, 2011.
- [34] A. A. Nasser, A. A. Alkhalaidi, M. N. Ali, M. Hankal, and Al-olofe M., "A Study on the impact of multiple methods of the data normalization on the result of SAW, WED and TOPSIS ordering in Healthcare Multi-attributes Decision Making Systems based on EW, ENTROPY, CRITIC and SVP weighting approaches," *Indian Journal of Science and Technology*, vol. 12, no. 4, pp. 1–21, Jan. 2019.
- [35] A. A. Nasser, A. A. Alkhalaidi, M. N. Ali, M. Hankal, and M. Al-olofe, "A Weighted Euclidean Distance - Statistical Variance Procedure based Approach for Improving The Healthcare Decision Making System In Yemen," *Indian Journal of Science and Technology*, vol. 12, no. 3, pp. 1–15, Jan. 2019.
- [36] А.А. Насер, "Концепция Построения Информационной Системы Вуза На Основе Структурно-Функционального Анализа Информационных Потокaв," *Вестник Апк Верхневолжья. № 1 (17), С. 81-85, 2012.*
- [37] А.А. Насер, "Насер Построение Сети Вуза На Основе Структурно-Функционального Анализа Информационных Потокaв," *Вести Высших Учебных Заведений Черноземья. № 4 (26). С. 56-60, 2011.*
- [38] А. А. Насер, А. А. Гуламов , "Информационная Модель Процессов Информационно-Аналитического Обеспечения В Вузе, Известия Юго-Западного Государственного Университета," № 5-1 (38). С. 61-64, 2011.
- [39] А А Насер, "Методы и алгоритмы интеллектуальной системы диагностики сосудистой патологии сетчатки глаза на основе контурного спектрального анализа и нейросетевого моделирования," *Юго-Западный государственный университет*, 2012.
- [40] Abed Saif Ahmed Alghawli, Adel A. Nasser, Mijahed N. Aljober, "A Fuzzy MCDM Approach For Structured Comparison of the Health Literacy level of Hospitals," *International journal of computer science and applications*, vol. 12, no. 7, pp. 81–97, 2021.
- [41] Adel A Nasser, Abdualmajed Al-Khalaidi and Mijahed N. Aljober, "Measuring the Information Security Maturity of Enterprises under Uncertainty Using Fuzzy AHP," *International Journal of Information Technology and Computer Science(IJTCS)*, 2018; 10, 10-25.
- [42] A. Nasser, "Information security gap analysis based on ISO 27001: 2013 standard: A case study of the Yemeni Academy for Graduate Studies, Sana'a, Yeme,n". *Int. J. Sci. Res. in Multidisciplinary Studies Vol*, 3(11).2017.
- [43] A. N. Al-Shameri, "Hierarchical Multilevel Information security gap analysis models based on ISO 27001: 2013," *International Journal of Scientific Research in Multidisciplinary Studies*, 3(11) , pp 14-23, 2017.
- [44] S. K. Manakandan, I. Rosnah, , R. J. Mohd, , and R. Priya, "Pesticide applicators questionnaire content validation: A fuzzy delphi method. ," *Med J Malaysia*, vol. 72, no. 4, pp. 228-235, 2017.
- [45] Yu-Lung Hsu, Cheng-Haw Lee, V.B. Kreng, "The application of Fuzzy Delphi Method and Fuzzy AHP in lubricant regenerative technology selection," *Expert Systems with Applications*, vol. 37, no. 1, pp. 419-425, 2010.
- [46] Rawabeh A. T., Haqiqi N., "The phenomenon of polygamy and its effects on relationships within the family," thesis in sociology organization and social dynamics, University of Algiers, 2011.
- [47] M. I. Aqqah, "The Social and Cultural Factors that Lead to The Phenomenon of Divorce in Light of The Social Changes in Palestinian Society: A Study in The Southern West Bank from 2013 to 2016", *Journal of the Faculty of Education - Assiut university*, vol. 35, no. 3, pp. 140–181, 2019.
- [48] N. Jain, "Forced Marriage as a Crime against Humanity: Problems of Definition and Prosecution," *Journal of International Criminal Justice*, vol. 6, no. 5, pp. 1013–1032, Nov. 2008.
- [49] A.-C. Zuntz, G. Palattiyil, A. Amawi, R. Al Akash, A. Nashwan, A. Al Majali, and H. Nair, "Early marriage and displacement a conversation: how Syrian daughters, mothers and mothers-in-law in Jordan understand marital decision-making," *Journal of the British Academy*, vol. 9, pp. 179–212, 2021.

- [50] J. C. Martín, P. Moreira, and C. Román, "A hybrid-fuzzy segmentation analysis of residents' perception towards tourism in Gran Canaria," *Tourism Economics*, vol. 26, no. 7, pp. 1282–1304, Sep. 2019.
- [51] J. Zhang, J. Cai, Z. He, C. Pu, and G. Tang, "Analysis on the Differences of Health Resources Allocation in Undeveloped Areas of Chongqing, China: A Cross-Sectional Study," *Journal of Service Science and Management*, vol. 13, no. 02, pp. 244–260, 2020.

A Penetration Testing on Malaysia Popular e-Wallets and m-Banking Apps

Md Arif Hassan*, Zarina Shukur, Masnizah Mohd

Center for Cyber Security, Faculty of Information Technology
National University Malaysia (UKM), 43600 UKM, Bangi, Selangor, Malaysia

Abstract—e-Wallets and m-banking apps became more and more popular in the developed world, approaching a point of tipping. This can be due to the global use of big and small merchants of paying equipment and the ubiquity of e-wallet and m-banking apps adoption. Many consumers are using e-wallets and m-banking apps that can be an effective cybercrime option. e-Wallets and m-banking apps allow financial transactions via smartphones that give cybercriminals a lucrative opportunity. Mobile technology has become increasingly mainstream and continually strengthening, with the focus on mobile apps protection and forensic analysis developing. In this paper, the security aspect of five popular e-wallets in Malaysia were analyzed. This paper also provides a security analysis of another five leading m-banking apps. The security analysis is based on a security principle that is recommended by Open Web Application Security (OWASP) under Mobile Security Testing Guide (MSTG) and Mobile Security Threats (MST). The static analysis has been done by using three mobile application-testing tools. This study included a variation of vulnerability scanning, code review and, most significantly, penetration testing. Each app complied with the security requirement, but their security features and characteristics, such as encryption, security protocols, and app services, are different to each other. This study was carried out using a DELL computer with Intel Core i7 CPU, 3.40 GHz CPU, 6 GB RAM. Finally, the results revealed the secure e-wallet and m-banking apps among the selected apps.

Keywords—Electronic payments; e-wallet; m-banking; android; static analysis; security analysis

I. INTRODUCTION

Nowadays, the development of technology advances has brought one of the pioneers of innovation in financial institutions. With the development of Fintech worldwide, there will still be enough challenges for those interested in adopting the technology. As part of their everyday transaction payment choice, several countries have already introduced the use of electronic payments. The payment methods used by consumers have great impacts on the future of the financial system and the business model of a country. Mobile payment services are increasingly popular in the banking world and are capable of replacing cash and becoming the most popular platform in the coming years. Fintech developments have changed payment systems in Malaysia. Malaysia has taken seriously the development of cashless societies in particular. The Central Bank of Malaysia intends to migrate towards a new cashless sector, in alignment with its financial plan 2020, with the intention of increasing efficiency in the financial sector [1]. For current stage, the most commonly used cashless payment methods are credit cards/debit cards, internet banking and

cheques [2]. e-Wallet appears to be a new trend of mobile payment in recent years. In effort to enhance the use of e-wallet in Malaysia, the e-wallet users included in the Malaysia Budget 2020 are granted an RM30 reward [3]. e-Wallet is a modern age of technologies that easily recognizes consumer interest, making our transactions very convenient and efficient [4-5]. Security is among the most crucial factors influencing consumers' determination to use e-wallet apps [6-7]. Cybercrime is a challenge to mobile payment systems, and is obviously not the only concern, although many consider that it is the greatest challenge in the field of mobile privacy. Many e-wallet application developers are financially motivated, and as a result, it is common to overcome challenges quickly in order to save time and money. Cybercrime possibilities are becoming more challenging, hence humans want to share their understanding of certain ways hackers could interfere with e-wallet apps. Those who hope this point of view will help people realize how cyber criminals think, because if everyone know that, then users will continue to defend ourselves by securing certain points of attack.

In addition to the design of the e-wallet apps implementation, it is necessary to preserve the protection and privacy of user data in general [8]. A safety intelligence survey found [9] that 400 leading companies, 40% of them don't even scan their code for security flaws. Besides all these apps, mobile money applications effectively cover high-level private financial and sensitive information, where protection needs to be of the extreme significance as vulnerabilities or risks cannot be tolerated, as absolute protection is necessary. This paper studies e-wallet products associated with e-money issuers listed in National Bank of Malaysia websites, by focusing on its mobile payment feature. There are 42 e-money certificates has issued by Central Bank of Malaysia, including 5 banks and 37 non-banks [10] and [11]. Security in general has become an increasing concern currently, particularly with the evolution of mobile payments, and the almost daily vulnerabilities found in operating systems and applications are what makes it more challenging [12]. This study conduct a penetrating testing using three static analysis tools which is recommended by Open Web Application Security (OWASP). The recommendation of OWSAP related to three static tools are show in Table I.

TABLE I. STATIC TOOLS SECURITY PRINCIPLE

Tools	Security principle	Suggested by
MobSF	OWSAP	Mobile Testing Guide
MARA	OWSAP	Mobile Security Threats
AndroBugs	OWSAP	Mobile Testing Guide

*Corresponding Author.

In this study, our main contributions are as follows:

- We perform a static analysis among five e-wallet and m-banking apps, specifically on security issues targeted for Android applications.
- In order to detect repackaging threats, we evaluate successful solutions and recognize the vulnerabilities.
- We discover the most secure bank and non-banks mobile application among selected application using static analysis tools.

This paper is bifurcated into four sections. Introducing the payments and its related study, which is already discussed above. The second section presented the five bank and non-bank issuers in Malaysia and its association with e-wallet products. Section 3 discussed the proposed methodology of the analysis. Section 4 presented the experimental result of the methods and Section 5, provides the discussion of the findings and finally, the conclusion b presented in Section 6.

II. LITERATURE REVIEW

Electronic payment systems have risen in popularity in the last 20 years because of their significant contribution in modern electronic transactions. Electronic transactions are a financial exchange between buyers and sellers available on the internet [13]. The payment system electronically originally referred to as a payment process through an electronic network [14] which a user may use to make online payments for products and services [15]. Among electronic payment system nowadays, e-wallet and m banking is one of the most famous payment system. The definition and their functionality is described in the next subsection.

A. e-Wallet and m-Banking

e-Wallet is the new invention of finance technology that make our transaction and payment easy and fast. e-Wallet is a virtual storage system [16] that can capture your identity and digital credentials and offer to an electronic gadget or online service that pro-vides a person to commit electronic purchase [17-18]. The e-wallet includes two elements, namely software and information. The software holds all the information contained in a wallet that encrypts confidential personal data. In comparison, the data are all information, such as customer ID, card information and shopping addresses, provided by the customer. There are quite variety of e-wallet services established worldwide.

For the past several years, mobile banking has grown in popularity across many segments of society. M banking is a subcategory of electronic banking that combines both the basics of banking and the distinct features of mobile payment [19]. M banking refers to the delivery and use of banking and financial services using mobile communications de-vices [20]. The range of services available might include the ability to conduct bank and stock market transactions, manage accounts, and obtain personalized data. Mobile banking, often known as m banking, is a terminology for using a mobile device such as a phone to perform balance checks, account transactions, payments, and credit applications. It is the convenient, simple, secure, anytime and everywhere in the world. The functionality

of the apps for each bank and non-bank issuers is listed in Table II and Table III.

- App based system: Application settings are also important since they allow user to personalize the program to his own needs. This will feature profile, payment, and security options, among other things.
- Fingerprint: Fingerprint identification is one of the most well-known and used bio-metric technologies. The fingerprint biometrics is useful and cost effective. Moreover, it can be quickly installed and used under any environmental conditions.
- Bills: This is an important e-wallet function because today's consumers like to pay all of their bills online, including utilities, mortgages, loans, rent, and tuition, to mention a few. e-Wallets are becoming a vital aspect of daily life as digital cash becomes more prevalent, and they should be able to give an easy bill payment option, whether it is a prepaid or postpaid payment service.
- Transfer of money: Transfer money between the payer and payee wallets in seconds rather than hours or days. This function has many advantages, including the ability to make payments at any time and from any location, the ability to make funds available immediately, and the ability to manage personal and business funds.
- Payment history: Any registered member will be able to view all the transaction de-tails in these features. After a successful login, the customer will be able to check order history.
- User account: To carry out a transaction from e-wallet to another e-wallet, the customer must be a user account. The user has been using a registration form. In order to register, the user must fill out all the fields required in the form. The user can access a variety of services using their user account.
- Pin Code: A personal identification number (PIN) is an unique code that must be in-put in order to complete certain banking transactions with a mode of payment. The purpose of a personal identification number (PIN) is to increase the level of security in electronic transactions.
- Top-up: By using these features, users could access the multiple bank list for top-up money. In these features, users need to choose how much they want to update. User chose their favorite bank account whenever. The payment gateway is submitted to the recipient of the application directly. The features help to users to add their balance into their e-wallet.
- QR code: In-store payments can be made using e-wallets using contactless methods such as near-field communication and Quick Response (QR) codes. QR Code is a form of 2D bar codes [21]. The QR code may be readily scanned with a smartphone camera [22]. The barcode is read by the smartphone, which then launches an associated apps or website.

- Open loop: Open loop mobile payment systems allow customers to pay from a centralized e-wallet at many different places. It is easy to comprehend closed payments as a gift card and open payments as a credit card [23].
- Add Money: Add money is used only for the logged-in users. It is connected to a payment gateway. The user can add or choose their bank account before transaction. User can add money with their registered bank account details or debit/credit cards using this function.
- Request money: Anyone may send his or her friend or family member a message to ask him or her to pay his or her money.
- Withdraw money: Users can transfer cash from their account to their connected bank account through the withdraw money functionality. Users may digitally withdraw money from the wallet into any bank account without the inconvenience of paper bills or currency, such as receiving money from sources, collecting money from distributors, in-store or online payments from consumers, or collecting money from sources.
- Voucher: Marketers and merchants are fully aware of the value of coupons and re-rewards. e-Wallets are a great platform for providing these benefits to value customers in a timely manner. As a result, features that make it simple to create and manage coupons, discounts, tickets, and loyalty points are essential for an e-wallet solution and may help e-wallet software stand out in the market.

In the following Table, II and III means- the apps have the properties, whether 0 means the apps do not have the properties. 0 means the apps do not have the properties whether 1 means the apps have the properties.

Table II and Table III revealed e-wallet apps functionality by bank and non-bank issuer where, most of them have same features except Non-Bank5. It shows that there are common and additional properties of e-wallet apps. There are 10 common e-wallet apps functionality by bank and non-bank issuer. They are pin code, fingerprint, bills, transfer of money, top-up, and payment history, add money, and voucher, Request money, and Withdraw money. Each e-wallet playing its own role and position in electronic payment system, several e-wallets such as Non-Bank1, Non-Bank3, and Non-Bank4 focusing on withdraw money from ATM booth, while others focus on online transaction.

B. The Open Web Application Security Project

The OWASP is a non-profit organization that works to improve software security. OWASP relies on an 'open community' concept, that enables anyone to participate in and assist to projects, events, online forums, and other services. Every three years, OWASP identifies the 10 most critical web application security risk types. This "top ten" list shows an agreement on the most serious security problems. The following were the top ten vulnerabilities as reported in OWASP 2017 are:

- 1) A1—Injection (SQL, OS, and LDAP).
- 2) A2—Broken Authentication.
- 3) A3—Cross-Site Scripting (XSS).
- 4) A4— Sensitive Data Exposure.
- 5) A5—Security Misconfigurations.
- 6) A6—Sensitive Data Exposure.
- 7) A7— Broken Access Control.
- 8) A8— Using Components with Known Vulnerabilities.
- 9) A9— Insufficient Logging and Monitoring.
- 10) A10— Insecure Deserialization.

The Open Web Application Security Project guide must be followed in some circumstances while developing web applications. The details of each vulnerability report can be found in [24].

TABLE II. THE FUNCTIONALITY OF BANK ISSUER M-BANKING APPS

Properties	Bank1	Bank2	Bank3	Bank4	Bank5
App based system	1	1	1	1	1
Fingerprint	1	1	1	1	1
Bills	1	1	1	1	1
Transfer of money	1	1	1	1	1
Payment history	1	1	1	1	1
PIN code	1	1	1	1	1
User account needed	1	1	1	1	1
top-up	1	1	1	1	1
QR code	1	1	1	1	1
Open loop	1	1	1	1	1
Voucher	1	1	1	1	1
Request money	1	1	1	1	1
Add money	1	1	1	1	1
Withdraw money	1	1	1	1	1

TABLE III. THE FUNCTIONALITY OF NON-BANK ISSUER E-WALLET APPS

Properties	Non-Bank1	Non-Bank2	Non-Bank3	Non-Bank4	Non-Bank5
App based system	1	1	1	1	1
Fingerprint	1	0	1	1	0
Bills	1	1	1	1	1
Transfer of money	1	1	1	1	0
Payment history	1	1	1	1	1
PIN code	1	1	1	1	0
User account needed	1	1	1	1	1
top-up	1	1	1	1	0
QR code	1	1	1	1	1
Open loop	1	1	1	1	1
Voucher	1	1	1	1	1
Request money	1	1	1	1	0
Add money	1	1	1	1	1
Withdraw money	0	1	1	1	1

C. Android Platform

Android is an open source smartphone operating system that was originally developed by Android Inc. and then acquired by Google with financial support. The first beta edition of Android was launched in November 2007 and the first stable version 1.0 followed in September 2008. Android is the world's most used mobile operating system, dominating the smartphone industry with an 82.8 percent share in 2015 [11]. That is good reason for this paper to perform the study on Android by itself. With the rising number of providers, each with its own Android OS version, the Android environment has been massively decentralized in recent years, ensuring that each has possible different vulnerabilities on top of some android platform problems, which is the total contrast on the IOS side where Apple maintains that a more compact closed ecosystem is accessible. IOS, though, is still suffering from smartphone protection concerns, and the android suffering is even higher. In comparison, being an open source based on Linux, Android makes it even easier to deal with, because it is much easy and popular for Android to remove the source code, scan the files, and include vulnerable code in applications.

D. Static Analysis

We chose static analysis as our vulnerability measurement technique despite reported vulnerabilities for several reasons. Unlike human code review or penetration testing, which results in reported vulnerabilities, static analysis is a purpose, repeatable, and scalable method for evaluating vulnerabilities. Static analysis tools employ a certain algorithms and rule sets every time and may scan a project in hours rather than days or weeks. Vulnerabilities can remain latent in code for years before a researcher discovers and re-ports them, thus the number of reported vulnerabilities is likely to be underestimated by an unspecified amount. Using static analysis, we could investigate the evolution of an application's vulnerabilities details. The number of vulnerabilities we uncovered with our static analysis technique far exceeded the number disclosed for the group of applications. Vulnerabilities of the same type in the same application version must be combined into a single item because of the Common Vulnerabilities and Exposures (CVE) criteria, which explain some differences in performance. Fig. 1 depicts the vulnerability static analysis process.

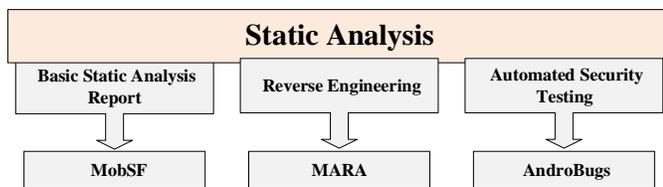


Fig. 1. Static Analysis Process.

1) *MobSF basic static analysis tools*: Mobile Security Framework that is an automated pen-testing framework capable of performing static and dynamic analysis. The framework also includes REST APIs that enable developers to use continuous integration and continuous delivery (CI/CD) pipelines to test their apps automatically with each build.

MobSF v2.0 is open-source and written in Python 3.7. It is released under the GNU Public License v3.0. The study's installed version was a Docker container provided and maintained by the authors of MobSF. This simplified the installation process and allowed the service to be dismantled or rebuilt on demand. Users can interact with MobSF's graphical user interface by navigating to localhost: 8000 when the Docker container has successfully started up. The security analysis of MobSF is depends on the following properties:

- **Signer Certificate**: A signing certificate encrypts an app, ensuring that the code underlying it is protected and that no one is defrauded.
- **Application Permission**: In Mobile application, several permissions that are classified as dangerous or acceptable. Understanding which permissions can lead to further damage is critical from the perspective of a security analyst. For example, if an application has access to external media and stores essential information on the external media, the files stored on the external media are globally accessible and writable, which might be harmful. Android app permissions can give apps control of users phone.
- **Network Security**: Details on network security issues relating to the application can be found in the network security section. These flaws might lead to critical attacks like man in the middle attack.
- **Android API**: The Android operating system gives a framework API for apps to interface with the Android underlying system.
- **Browsable Activities**: Browsable Activities can control how the device should react when the user clicks on a link in the web browser.
- **Security Analysis**: Details about security issues relating to the application can be found in security. These problems can lead to critical attacks. The security analysis is including Network Security, Manifest Analysis, Code Analysis, Binary Analysis, NIAP Analysis and File Analysis.
- **Manifest Analysis**: Manifest Analysis may extract a variety of data from an Android manifest file, such as which activities are exported, if the app is debug gable, data schemas.
- **Code analysis**: The code analysis part of the MobSF tool is one of the more interesting aspects. MobSF has analyses, evaluated parts of the application's behavior using industry security standard practices such as OWASP Mobile Security Testing Guide (MSTG), and mapped the vulnerabilities using OWASP Top 10. Furthermore, Common Weakness Enumeration and Common Vulnerability Scoring System scores (CVSS) are stated, which might be helpful in different analyst scenarios and make the development of reports a bit more straightforward for developers and analysts.

- **Malware Analysis:** Malware analysis is the domain malware check. MobSF extracts the hard-coding or application-using URLs/IP addresses, presents the malware status, and uses the ip2location to indicate its Geo location. APKiD is used to identify different packers, compilers, and hypocrites.
- **Reconnaissance:** Reconnaissance is used to identify the applications URLs, firebase DB, emails, trackers, and string and hard-coded secrets. This is all done using the decompiled source code.
- **Components:** Components are used to identify the details information regarding activities services. This summarizes the android APK skeleton. The components are including Activities, Services, Receivers, Providers, Libraries and Files.

In contrast to traditional desktop and web apps, mobile applications have unique security challenges. With mobile app security, the MobSF tool is widely employed. According to MobSFs, security score, CVSS and trackers' detection determine the outcome.

- **Security Score:** Security scoring is one of the important features to calculate the overall result. This security scoring is based on, if it introduced an issue to an app that already has higher average CVSS than that issue, app would actually have higher score than before even though it now has more issues. The developer improves the tool's security score. Since the average CVSS score is used, an app with one major issue and several minor ones may score higher than one with only one major issue [25]. Currently, app score is calculated as:

$$\text{avg_cvss} = \text{round}(\text{sum}(\text{cvss_scores}) / \text{len}(\text{cvss_scores}), 1)$$

$$\text{app_score} = \text{int}((10 - \text{avg_cvss}) * 10)$$

- **CVSS:** The CVSS score can be utilized to determine the severity of vulnerabilities found in apps. The CVSS Calculator can be used to calculate the CVSS score. The formulas given in the CVSS specification are used in the calculation [26]. The CVSS Metric Values are shown in Table IV. The details CVSS scores provided by MobSF can be seen in supplementary file (Table I-III) [26-27].
- **Tracker Detection:** Each app may make use of third-party trackers. MobSF analyses the detected tracker in the system's APK using the open source Exodus-Privacy web tool. Tracker analysis can be found in two ways, such as crash reporters and analytics. Crash reporters are those that look into crashes that happen when the program is running normally. Alongside, analytics tracker collects information on how users interact with the app, such as how much time they spend in it, which features they utilize [28].

2) *Reverse engineering:* The method of extracting technology or design information from something man-made is known as reverse engineering [29-30]. The theory behind, it has for centuries been understood that destroying something would help you understand it, evaluate it and even twist it to

achieve another task. In the computer security industry, reverse engineering is commonly used to analyze and exploit viruses and malware, vulnerability detection, binary code auditing, and development [30-31]. In this paper, reverse engineering was introduced after automated security tested as a second part of the static vulnerability analysis.

Reverse engineering is in particular used to see how engineers have constructed this specific system and how they perform essential protective activities and specifications [32]. These are some examples of what we can look for in Android security tests while using reverse engineering: database link, DB name, or DB password, certain hard-coded usernames or passwords that is used to accessing the database [33]. The application's APIs to see whether any of them are compromised, or the API key, and to check for a known vulnerable method after downloading the source code. This section will discuss the tools used in Reverse Engineering, and then the procedure followed to identify vulnerabilities in the selected applications, and will finally go over the results obtained from reverse engineering. During the reverse engineering, the method was initially focused exclusively on the Mobile Reverse Engineering & Analysis framework (MARA) which takes an APK file and delivers the source code in a language that is easily understood in Smali. Fig. 2 indicates the process that follows an application for reverse engineering.

TABLE IV. CVSS METRIC VALUES

No	Metric	Metric Value	Description
1	Attack Vector	Network Adjacent Network Local Physical	The attack vector defines the circumstance in which a vulnerability can be exposed.
2	Attack Complexity	Low High	This metric specifies the conditions that must exist outside the attacker's control in order to exploit the vulnerability. Depends on the situation, unique conditions that need a measurable amount of preparation or execution are necessary for the exploitation of the vulnerability, the metric's score might be low (L) or high (H).
3	Privilege Required	None Low High	The level of privilege necessary to exploit the vulnerability is defined by this metric. Its value is None (N) if the attacker does not need any permission; Low (L) if the attacker just needs basic privileges to change user-owned settings and files; or High (H) if the attacker needs major privileges to affect component-wide configurations and files.
4	User Interaction	None Required	This metric can have the values None (N) or Required (R) depending on whether the vulnerability is exploited with or without user participation.
5	ImpactConf, ImpactInteg, ImpactAvail	High Low None	This metric are refer to the Confidentiality Impact, integrity and availability impact.

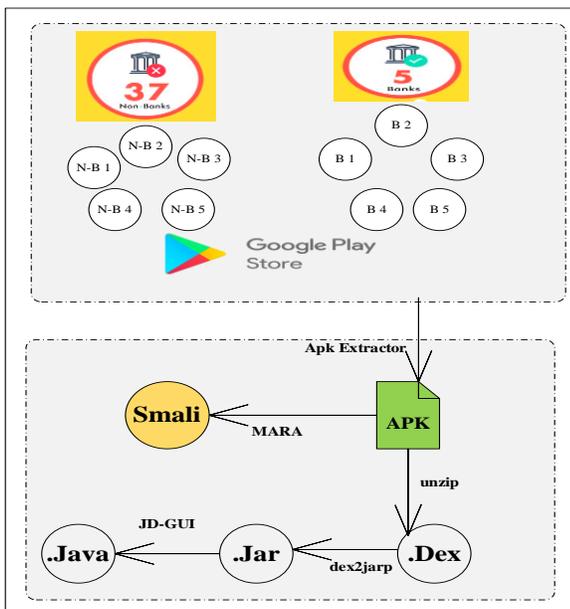


Fig. 2. Reverse Engineering Process [29].

The Mobile Application Reverse Engineering and Analysis (MARA) framework was the technology utilized to execute penetration testing. MARA brings together commonly used reverse engineering and analysis tools for mobile applications to test mobile apps against OWASP mobile security threats and vulnerabilities. MARA is a bash script that combines several prominent android reverse engineering and vulnerability analysis tools everything into solution. The goal was to make the workflow easy to utilize for security researchers and developers. MARA can do dynamic and static analysis requiring no additional post-installation configuration. MARA does not have a graphical user interface; it was created only for terminals. Most of the sub-tools are developed in various Python versions. MARA is evaluated based on six security attributes, which are APK Reverse Engineering, APK Deobfuscation, APK Analysis, APK Manifest Analysis, Domain Analysis, and Security Analysis. Each of the following six security attributes can be found in details [34].

Critical, High, Medium, Low, Info, and Detection issue are the different levels of se-verity [35]. MARA reversed all the chosen applications and retrieved the complete source code. The reverse results obtained by checking the source code individually are provided in Section 4.

3) *Automotive testing*: This segment would present the automated security checks carried out in the chosen banking and non-bank payment methods, processes and outcomes. Automated protection testing is an essential part of the paper's static vulnerability review, which is valuable since it provides a summary of where vulnerabilities can be found in the application. This chapter starts with a summary on chosen tools, and then explains how they are used then shows the results of automatic safety monitoring. All the apps checked for AndroBugs security have received an initial description of where to check for vulnerabilities, while the findings presented by AndroBugs seemed to provide a clear

understanding of the system build and possible vulnerabilities in most instances.

In November 2015, Yu-Cheng Lin released the AndroBugs framework, an open source vulnerability scanner for Android apps. AndroBugs is a Python-based static analysis tool that analyzes for common vulnerabilities in Android apps, it also checks the code is missing best practices and checks dangerous shell commands [29]. AndroBugs seemed to provide a clear understanding of the system build and possible vulnerabilities in most instances. The details of calculated CVSS for AndroBugs can be seen in supplementary file (Table IV) [26]. The possible vulnerabilities can be found in all apps were the following:

- **Runtime Command**: This is because AndroBugs establish in the code a serious function "Runtime.getRuntime ().exec (...)". This feature allows a user to enter the shell and then modify the commands within it.
- **Fragment Vulnerability**: Since AndroBugs found a 'Fragment' or 'ActionBarSherlock Fragment' in the software, which was vulnerable to Android before 4.4 on phones. Any intruder who runs a code capable of eroding the Android Sandbox, which means access to confidential information not accessible by the application, is vulnerable to using this application on an old Android device.
- **SSL Certificate Verification**: However, this application does not validate the SSL certificate validation, so it causes the self-signed Common Name (CN) certificates for Secure Sockets Layer (SSL) to be expired or to be unacceptable. This is undoubtedly a vital weakness, since it enables attackers to carry out Man in the middle (MITM) attacks.
- **SSL Implementation**: That ensures that such a self-defined application will accept all common names as "HOSTNAME VERIFIER." This allows any attacker with a valid certificate to carry out MITM attacks.
- **Implicit Service**: In other words, this application uses an implicit decision to conduct a service, which is dangerous, since the answering service is not identifiable and the user cannot see which service.
- **Web View Vulnerability**: This implies that AndroBugs find the "addJavaScriptInter-face" method in an application code, which a weakness that JavaScript may use in devices is running android before 4.2 to manage the application.
- **Android Manifest**: This shows that this app has high privileges for AndroBugs. An-droBugs find that the "Mount Unmount FileSystems" android permission is included in this program, which has not been justified as the permission authorization permits removable mounting and mounting of file systems and the Android developer website notes that the application is not used by third party applications. The application is not mounted.

- **Key Store Protection:** AndroBugs therefore find that this application does not adequately secure its Key Store because it appears that it uses byte array and SSL pinning using a hard-coded certificate information. The details of the reverse engineering process are mentioned in Section 2.4, respectively. The existing work and their finding is shown in Table V.

TABLE V. THE EXISTING WORK AND THEIR FINDING

Authors	Method	Country	Finding
Reaves et al. 2015 [39]	Manual analysis	USA	This article completed manual analysis on 7 Android m-banking apps. These apps were tested for SSL/TLS bugs, cryptography, and identity leakage and access control. The findings confirm that the majority of these apps fail to provide the protections needed by financial services.
Filiol and Irolla, 2015 [37]	Static and Dynamic analysis	France	This study executed static and dynamic analysis on 50 Android m-banking apps
Zheng et al. 2017 [38]	Repackaging Attack	Australia	This article examine common security attacks on mobile apps, whether they are performing preliminary tests to determine the effectiveness and complexity of mobile device security attacks using repackaging attacks to obtain victim information.
Chothia et al. 2017 [36]	TLS testing methods	UK	This article presents a security analysis of the 15 m-banking apps issued by leading UK banks. The primary goal was to find the bugs in these apps' TLS implementations.
Chanajitt et al. 2018	Forensic analysis	Thailand	This articles focus on seven Android m-banking apps in Thailand. Several of the applications examined do not perform root device identification, do not encrypted user data, or may be modified and installed as repackaged apps.
Bassolé et al. 2019 [40]	Vulnerability assessments	Africa	This article analyzed the vulnerability of mobile banking and payment applications on Android platforms. This article undertakes vulnerability assessments, allowing for a more informed analysis of the information security and privacy threats that African mobile banking and payment applications face. They specially evaluate login credentials and code vulnerability of these apps in particular to assess the risks of attacks connected to privacy and data confidentiality.
Yang et al. 2019 [41]	Comprehensive analysis	China	This article examines the existing third-party mobile payment ecosystem and identifies possible security concerns by doing an in-depth assessment against China, the world's largest mobile payment market. Aside from that, this

			article also uncovers seven incidences of security rule violations on the Android and IOS platforms.
Verderame et al. 2020 [45]	Static and Dynamic Analysis	Italy	This article describes a unique methodology based on a successful mix of static analysis, dynamic analysis, and machine learning techniques for determining whether a particular app either) has a Google Play privacy policy and ii) accesses privacy. This article also involves examining the compliance of third-party libraries that are incorporated in existing applications.
Majeti et al. 2021 [46]	Cryptographic primitives	India	This article looks at how cryptographic primitives are used in Indian mobile finance apps. They chose 36 apps from three distinct categories and evaluated the flaws separately.
Our study	Static analysis	Malaysia	To perform static analysis of 5 m-banking and non-bank e-wallet apps. The static analysis has been done by using three mobile application-testing tools that is recommended by OWASP.

III. METHODOLOGY OF PROPOSED STUDY

This section presents the methods, processes and results of the automated security tests on the applications pre-selected. The automated safety testing is part of the research paper static vulnerability analysis, with an overview of where vulnerabilities can be found. The analysis is important. This chapter starts with a briefing on selected the analysis tools, and then demonstrates the method and results of the automated security testing. Some of the previous study conducted static or dynamic analyzes among various country leading m-banking apps. The focus of this study is to analyze popular Malaysian e-wallet apps and m-banking apps to identify the security.

A. Information Gathering and Setup

Mobile testing tools can assist organizations in automating Android and iOS testing. The software for mobile application testing can minimize the time required for the test and the probability of human mistakes during testing. Varieties of technologies are available for testers to automate their test scripts nowadays. For the success of the objective, it is essential to choose the right path for particular apps. Various systems may have various risks. The key problem, for example, would be diversion of funds in a bank application.

For object testing, authors utilized a DELL machine with an Intel Core i7 CPU, 3.40 GHz CPU, and 6 GB RAM. The operating system is Windows 10 professional. A virtual machine has been installed name as VMware to create dual boot in the computer. The testing involved the process of first installing Kali Linux, Operating System (OS).

B. Analysis Process and Testing Object

Intruders do the test to evaluate if there are any flaws or weaknesses that can allow the penetration and exploitations during its operation [41]. The e-wallet applications for Android

Redmi 8 are executed to start research, to check if they are running without any error. The study included a variation of vulnerability scanning, code review and, most significantly, penetration testing. AndroBugs is used to automate security-testing tool, where MobSF, used for basic static analysis report. Finally, MARA tools used for reverse engineering checking. In checking mobile apps against the OWASP, MARA builds widely used reverse engineering and research techniques to test mobile applications [42].

The object contains e-wallet applications from leading and growing banks in Malaysia. The platform of Google Play Store, the official site for Android-based smartphone apps downloaded e-wallet applications on the mobile phone. The applications were then transmitted via a universal serial bus serial interface to the computer. A folder was then created with Malaysia e-wallet name, which was then dumped in the e-wallet APK for-mat. APK extension dumped on the desktop. All the selected applications have been security tested by Mobile Security Framework (MobSF), MARA and AndroBugs tool an immediate idea of where weaknesses might be found. The tools findings seem to provide a clear understanding of how the programs are designed and where it could lead to potential vulnerabilities. It was beneficial as a guide for static vulnerability analysis and to know where the weakness available. Fig. 3 shows the selected dataset of e-wallet programs with the analysis process.

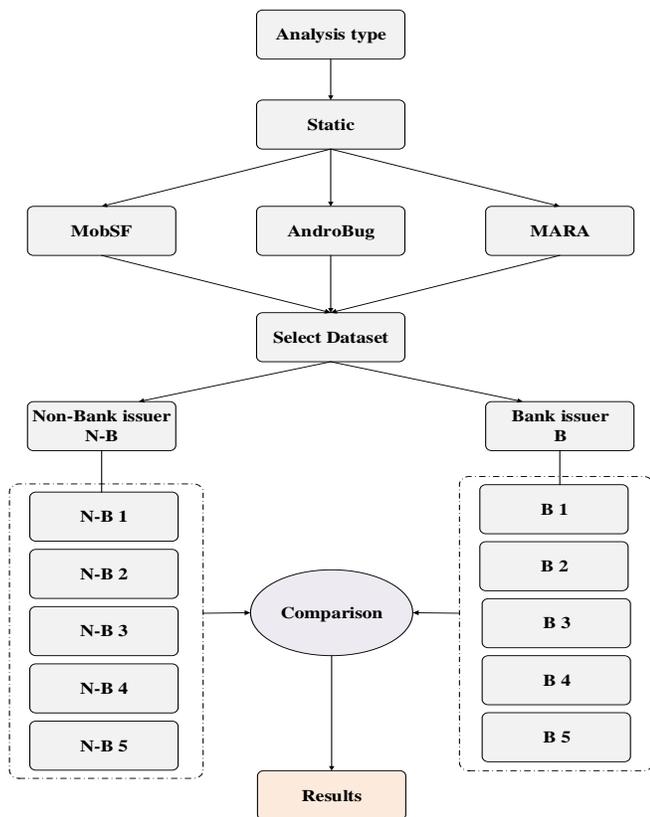


Fig. 3. Selected Testing Object for Analysis.

IV. ANALYSIS RESULT

After installation of the OS, updates and patches for the operating system were then installed from the Linux public repository to update the libraries that used as prerequisites for its installation and operation. The applications were then transferred to the computer via a universal serial bus data cable. The applications were then analyzed one by one tools and reports created and dumped in their respective application folders. The result of non-banks e-wallet apps using MobSF is showing Table VI.

Table VI shows the comparison result of select non-bank e-wallet apps. The analysis report is divided into three categories, such as security score, average Common Vulnerability Scoring System (CVSS) and tracker detection. The security score refers to the overall security results where the CVSS is an open framework for interactive the characteristics and severity of software vulnerabilities. Finally, tracker detection vulnerability checks and evaluates IT network and any device linked to it against thousands of Network Vulnerability Tests (NVTs) [43]. The most active security score in the analysis report is N-B4, which was observed 45 percent of the time, with an average CVSS of 7.0 and 8-tracker detection, which was the top security score APK. The second highest positions security score obtain from N-B3 with 40%, which CVSS rate average 6.4 with 6-tracker detection. From N-B1, N-B2 and N-B5, the same security score has been identified which 10% respectively. Nevertheless, in point of view their CVSS and tracker detection are not similar to their security score. The average CVSS of N-B1 and N-B2 are similar to 6.5 whether N-B5 is 6.9. The tracker detection rates are 6/323, 7/323, 3/323 and 2/323.

Table VII shows the comparison result of select bank issuer mobile apps. The most active security score in analysis report is, B2 with 85%, which average 7.5 CVSS with 0-tracker detection, which was the top security score APK. The second highest positions security score obtain from B1, B3, and B4, with same security score which 10% respectively. Nevertheless, in point of view their CVSS and tracker detection are not similar like their security score. The average CVSS of B1 with 6.7, B3 is 6.6, and B4 is 6.1. Whether the tracker detection rates are 2/319, 3/323, and 3/319. From B5 authors could not find the security score but except Average CVSS 6.8 and Trackers Detection 8/319, respectively. The result of banking apps using MobSF is shown in Table VII. The analysis report of non-banking e-wallet apps is shown in Table VIII.

TABLE VI. RESULT OF NON-BANKS E-WALLET APPS USING MOBSSF

Wallet name	Security Score	Average CVSS	Trackers Detection
N-B1	10/100	6.5	6/323
N-B2	10/100	6.5	7/323
N-B3	40/100	6.4	6/323
N-B4	45/100	7.0	8/323
N-B5	10/100	6.9	2/323

TABLE VII. RESULT OF BANKS APPS USING MOBSEF

Wallet name	Security Score	Average CVSS	Trackers Detection
B1	10/100	6.7	2/319
B2	85/100	7.5	0/319
B3	10/100	6.6	3/323
B4	10/100	6.1	3/319
B5	-	6.8	8/319

TABLE VIII. REPORT OF NON-BANKS APPS USING MARA

Wallet name	Critical	High	Medium	Low	Info	Detection Issue
N-B1	0	-	1	-	-	-
N-B2	0	-	1	-	-	-
N-B3	0	4	2	2	11	19
N-B4	0	6	1	2	11	20
N-B5	0	2	1	2	10	15

Table VIII shows each of the e-wallet application has 0 critical issues. From N-B4 with total 20 detection issues 6 high, 1 medium, 2 low and 11 info analysis report. In N-B3 total 19 detection issues has been collected where 4 high, 2 medium, 2 low and 11 info analysis report which the second highest. The third positions is N-B5 with total 15 detection issues where 2 high, 1 medium, 2 low and 10 info analysis report. N-B1 and N-B2 there is no critical issue but due to system trouble-shoot authors could not get the exact information of both wallet. The result of banking apps using MARA tool is shown in Table IX.

Table IX shows the comparison result of select bank issuer mobile Apps. Each of the banking application has 0 critical issues. From B1 total 19 detection is-sues has been collected where 4 high, 2 medium, 0 low and 11 info analysis report. From B3 and B5 collected a similar value total 17 detection issues where 3 high, 2 medium, 2 low and 10 info analysis report which the second highest, respectively. The third positions is B5 with a total 17 detection issues where 3 high, 2 medium, 2 low and 10 info analysis report. From B4 Apps there is no critical issue but due to system troubleshoot authors could not get the exact information. Finally, B1 has 0 critical issues with high threats, 2 medium and low threats along with 10 info are the most secure application compared to others. The result of banks and non-banks reports using MARA AndroBugs are shown in Table X and Table XI.

TABLE IX. REPORT OF BANKS APPS USING MARA

Wallet name	Critical	High	Medium	Low	Info	Detection Issue
B1	0	4	2	0	11	19
B2	0	0	2	0	6	8
B3	0	3	2	2	10	17
B4	-	-	-	-	-	-
B5	0	3	2	2	10	17

Table X and Table XI shows below, the comparison result of select bank and non-issuer mobile apps using AndroBugs. Each of the application has different kind of critical issue. The analysis report has been categorized into two part parts. From B4 with total 6 issues which is the most critical issue found in the bank issuer analysis report. From B1 and B5 authors collected similar value total 5 issues which the second highest, respectively. The third position is B3 with 3 issues.

TABLE X. RESULT OF BANKS REPORT USING ANDROBUGS

S/n	Properties	B1	B2	B3	B4	B5
1	Runtime Command Checking	No	No	No	No	Yes
2	Base64 String Encryption	No	No	No	No	No
3	SSL Security	No	No	No	No	No
4	Key Store	No	No	No	Yes	Yes
5	Implicit Intent	Yes	No	Yes	Yes	Yes
6	SSL Implementation Checking	Yes	No	No	Yes	No
7	SSL Connection Checking	Yes	No	Yes	No	Yes
8	SSL Certificate Verification Checking	No	No	No	Yes	No
9	<Web View>/Remote Code Execution	Yes	Yes	Yes	Yes	Yes
10	Fragment Vulnerability Checking	Yes	No	No	No	No
11	Android Manifest Critical Use Permission Checking	No	Yes	No	Yes	No

From B2 only single issues have been collected, which is the most secure using AndroBugs analysis. From non-bank, issuer author's collected total 9 issues from N-B5 which is the most critical issue found in the analysis report. The second highest positions are N-B2, with total 7 issues. From N-B3 and N-B1, similar value has been collected which are total 3 issues respectively. From N-B4, only single issues have been collected, which is the most secure using AndroBugs analysis.

V. DISCUSSION

Mobile apps are growing increasingly, with more consumers being able to access different forms of Android applications availability of a wide range of open Android markets. However, mobile apps threats are developing especially targeted towards mission critical mobile bank applications [44]. This study first analyze five types of bank and nonbank issuer e-wallet products, and then focus on the vulnerabilities and security issues based on static analysis. We evaluate the flaws in protection, critical security, Average CVSS against Malaysian 5 banking and non-banking e-wallet products publicly available. In this section, will present and analyze the outcomes of a data set security evaluation.

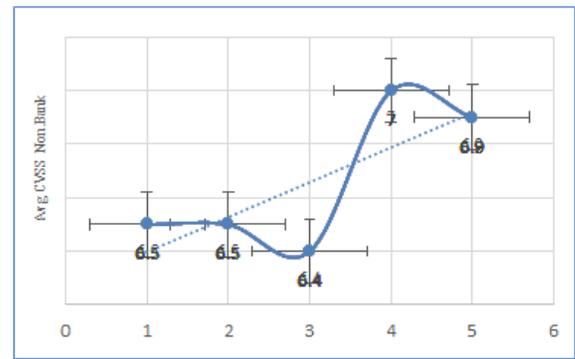
TABLE XI. RESULT OF NON-BANKS REPORT USING ANDROBUGS

S/n	Properties	N-B1	N-B2	N-B3	N-B4	N-B5
1	Runtime Command Checking	No	Yes	No	No	Yes
2	Base64 String Encryption	No	Yes	No	No	Yes
3	SSL Security	No	No	No	No	Yes
4	Key Store	No	No	No	No	Yes
5	Implicit Intent	Yes	Yes	Yes	No	Yes
6	SSL Implementation Checking	No	No	No	No	Yes
7	SSL Connection Checking	Yes	Yes	Yes	No	Yes
8	SSL Certificate Verification Checking	No	Yes	No	No	Yes
9	<Web View>/Remote Code Execution	Yes	Yes	Yes	Yes	Yes
10	Fragment Vulnerability Checking	No	No	No	No	No
11	Android Manifest Critical Use Permission Checking	No	Yes	No	No	No

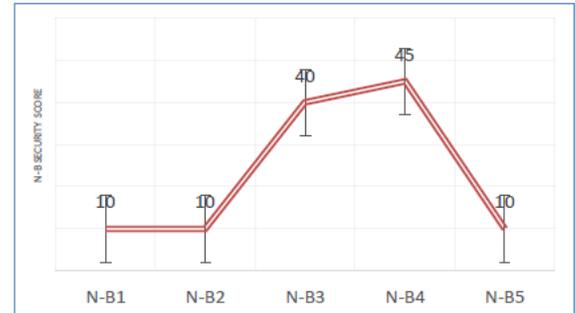
In Fig. 4 presents, the summary of MobSF results of the security tests performed on bank and non-banks e-wallet apps. The most active security score in the analysis report is N-B4, is 45%, which averages 7.0 Common Vulnerability Scoring System (CVSS) with 8-tracker detection, which was the top security score application.

It is clear from the results that N-B4 is quite secure related to the other applications. Table VIII presented the result of banks' apps using MobSF. The most active security score in the analysis report is B2, from which noticed 85%, which average 7.5 CVSS with 0-tracker detection, which was the top security score APK. After the analyzing of Table VIII, found that, N-B5 application have total 15 detection issues where 2 high, 1 medium, 2 low and 01 info analysis report which is quite secure related to the other applications is shown in Fig. 5(a) and Fig. 5(b) shown the high security banking where B3 and B5 are same result, respectively.

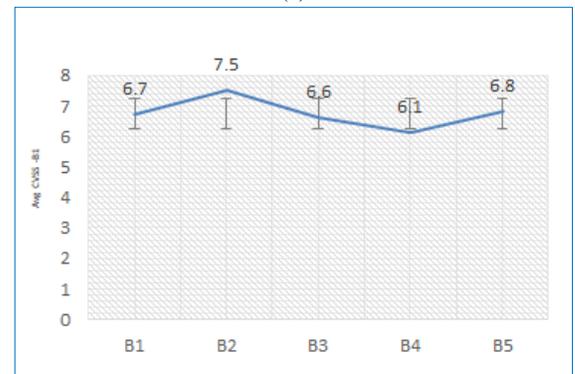
On the other hand, the low security low security apps of banking m-apps using MARA tool is B1, which is shown in Fig. 5 (b). Table IX shows the true seeing a report of banks' apps using MARA. Each of the banking application has 0 critical issues. However, the B3 and B5 have 0 critical issue with 3 high threats, 2 medium and low threats along 10 info and 27 detection issues which is the most secure application compared to others. Fig. 5(c) and (d) shown the low security bank and non-bank apps, respectively.



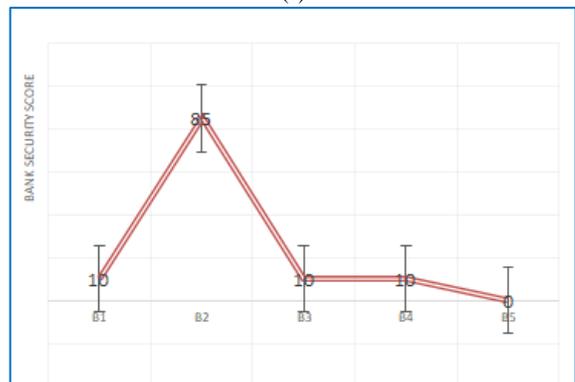
(a)



(b)



(c)



(d)

Fig. 4. (a) Average CVSS of Non-banks e-wallet Apps using MobSF Tool and (b) Non-bank e-wallet Apps Security Score using MobSF Tool and (c) Average CVSS of Banks m-banking Apps using MobSF Tool and (d) Bank m-Banking Apps Security Score using MobSF Tool.

VI. CONCLUSION

Mobile payment applications are very convenient, but the problem is that most mobile payment apps are not exactly appropriate. Companies and developers would need to limit the addition of features and services demand and continue protecting the apps. However, as explained in this paper, it is quite an impossible task to protect the apps, although nothing is 100% secured, but developers at least might make it much more difficult for hackers. This paper covers a security assessment of five non-bank e-wallet apps and five leading banks apps in Malaysian market. The Authors performed a static analysis on three pen-extraction mobile application-testing tools and compared with the Android application among them. The analysis notice that every apps have followed the security standard, but their security features and properties are different in point of view of how their customer demand. Finally, the most secure e-wallet and m-banking apps according to the three different tools, based on their security metrics, have been identified which is not our opinion. This study aims to increase research efforts on the progress of e-wallets and m-banking in Malaysia.

ACKNOWLEDGMENT

This research was funded by the Malaysia Ministry of Education, Universiti Kebangsaan Malaysia, under grants, KKP 2020/UKM-UKM/4/3 and FRGS/1/2021/ICT02/UKM/02/1.

REFERENCES

- [1] F. Nizam, H. J. Hwang, and N. Valaei, *Measuring the Effectiveness of E-Wallet in Malaysia*, vol. 786. Springer International Publishing, 2019.
- [2] 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.
- [3] L. T. H. Teoh Teng Tenk, Melissa, Hoo Chin Yew, "E-wallet Adoption: A Case in Malaysia," *Int. J. Res. Commer. It Manag.*, vol. 2, no. 4, pp. 135–3, 2020.
- [4] A. Hassan, Z. Shukur, and M. K. and A. S. A.-K. Hasan, "A Review on Electronic Payments Security," *Symmetry (Basel)*, vol. 12, no. 8, p. 24, 2020.
- [5] M. Salah Uddin and A. Yesmin Akhi, "E-Wallet System for Bangladesh an Electronic Payment System," *Int. J. Model. Optim.*, vol. 4, no. 3, pp. 216–219, 2014.
- [6] S. Z. Jesús Téllez Isaac, "Secure Mobile Payment Systems," *J. Enterp. Inf. Manag.*, vol. 22, no. 3, pp. 317–345, 2014.
- [7] M. A. Hassan and Z. Shukur, "Review of Digital Wallet Requirements," 2019 *Int. Conf. Cybersecurity, ICoCSec 2019*, pp. 43–48, 2019.
- [8] R. Kaur, Y. Li, J. Iqbal, H. Gonzalez, and N. Stakhanova, "A Security Assessment of HCE-NFC Enabled E-Wallet Banking Android Apps," *Proc. - Int. Comput. Softw. Appl. Conf.*, vol. 2, pp. 492–497, 2018.
- [9] IBM, "IBM Sponsored Study Finds Mobile App Developers Not Investing in Security." [Online]. Available: <https://www-03.ibm.com/press/us/en/pressrelease/46360.wss>. [Accessed: 15-Nov-2020].
- [10] EcInsider, "The e-wallet infinity war in Malaysia - Everything you need to know about e-wallet starts here." [Online]. Available: <https://www.ecinsider.my/2018/12/malaysia-ewallet-battle-landscape-analysis.html>. [Accessed: 01-Nov-2019].
- [11] A. Hassan, Z. Shukur, and M. K. Hasan, "Electronic Wallet Payment System in Malaysia," *Data Anal. Manag.*, vol. 54, pp. 711–736, 2021.
- [12] Y. Wang et al., "Identifying vulnerabilities of SSL/TLS certificate verification in Android apps with static and dynamic analysis," *J. Syst. Softw.*, vol. 167, 2020.

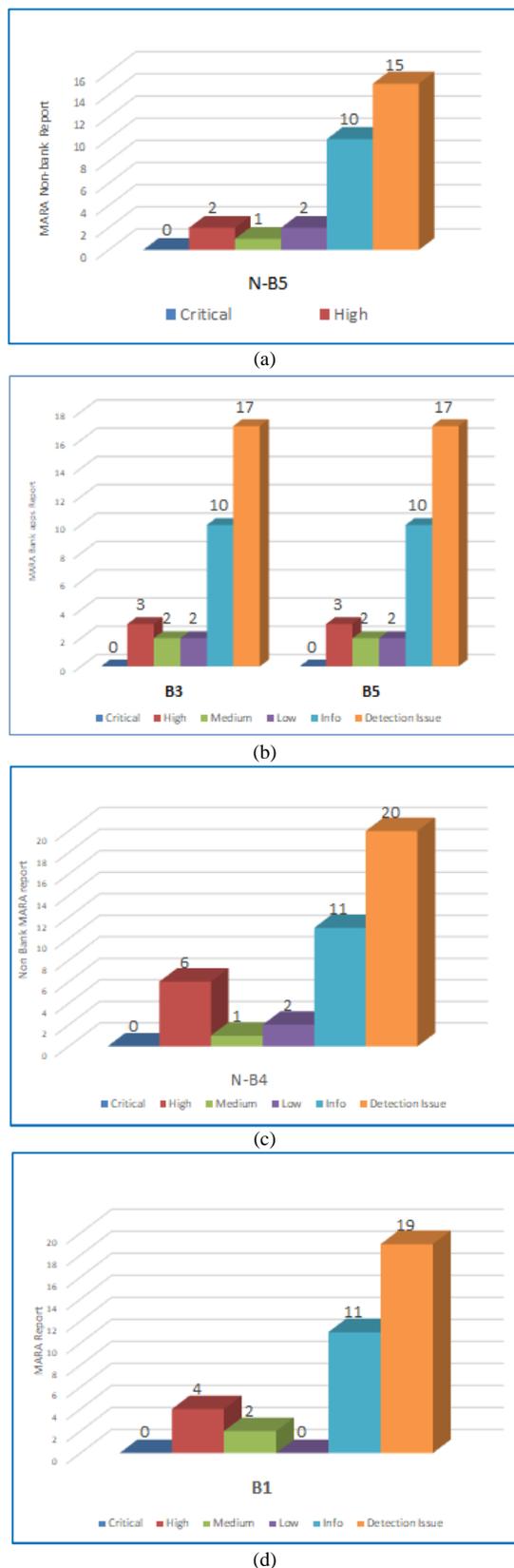


Fig. 5. (a) High security e-wallet apps using MARA tool and (b) High security banking apps using MARA tool and (c) Low security e-wallet apps using MARA tool and (d) Low security apps of m-apps using MARA tool.

- [13] P. Aigbe and J. Akpojaro, "Analysis of Security Issues in Electronic Payment Systems," *Int. J. Comput. Appl.*, vol. 108, no. 10, pp. 10–14, 2014.
- [14] M. A. Kabir, S. Z. Saidin, and A. Ahmi, "Adoption of e-payment systems: a review of literature," *Proc. Int. Conf. E-Commerce*, no. May 2016, pp. 112–120, 2015.
- [15] Rancho and P. Singh, "Issues and Challenges of Electronic Payment Systems," *Int. J. Res. Manag. Pharmacy(IJRMP)*, vol. 2, no. 9, pp. 25–30, 2013.
- [16] R. Batra and N. Kalra, "Are Digital Wallets the New Currency?," 2016.
- [17] M. Olsen, J. Hedman, and R. Vatrapu, "E-wallet properties," *Proc. - 2011 10th Int. Conf. Mob. Business, ICMB 2011*, pp. 158–165, 2011.
- [18] M. A. Hassan Z. Shukur, M. K. Hasan "An efficient secure electronic payment system for e-commerce," *Computers.*, 9(3), 66, 2020.
- [19] R. Safeena, H. Date, A. Kammani, and N. Hundewale, "Technology Adoption and Indian Consumers: Study on Mobile Banking," *Int. J. Comput. Theory Eng.*, no. December, pp. 1020–1024, 2012.
- [20] S. Shaju and V. Panchami, "BISC authentication algorithm: An efficient new authentication algorithm using three factor authentication for mobile banking," *Proc. 2016 Online Int. Conf. Green Eng. Technol. IC-GET 2016*, pp. 1–5, 2017.
- [21] J. Juremi, "A Secure Integrated E-Wallet Mobile Application For Education Institution," *Int. Conf. cyber Relig.*, 2021.
- [22] J. Zhang and Y. Luximon, "A quantitative diary study of perceptions of security in mobile payment transactions," *Behav. Inf. Technol.*, vol. 0, no. 0, pp. 1–24, 2020.
- [23] M. H. Sherif, *Protocols for Electronic Commerce*, vol. 53, no. 9, 2016.
- [24] OWASP, "Owasp-Asvs," no. October, 2015.
- [25] A. Abraham, "Improve security scoring of apps." 2020.
- [26] A. Maharjan, "Ranking of android apps based on security evidences," no. December, 2020.
- [27] A. Abraham and S. Dominik, "Mobile Security Framework." 2019.
- [28] V. Kouliaridis, G. Kambourakis, E. Chatzoglou, D. Geneiatakis, and H. Wang, "Dissecting contact tracing apps in the Android platform," *PLoS One*, vol. 16, no. 5 May, pp. 1–28, 2021.
- [29] H. Darvish and M. Husain, "Security Analysis of Mobile Money Applications on Android," *Proc. - 2018 IEEE Int. Conf. Big Data, Big Data 2018*, pp. 3072–3078, 2019.
- [30] PCI Security Standards Council LLC., "PCI Mobile Payment Acceptance Security Guidelines for developers," *Pci Dss Inf. Suppl.*, no. February, pp. 0–27, 2013.
- [31] Enisa, Security of Mobile Payments and Digital Wallets, no. December. European Union Agency for Network and Information Security (ENISA), 2016.
- [32] T. McDonnell, B. Ray, and M. Kim, "An empirical study of API stability and adoption in the android ecosystem," *IEEE Int. Conf. Softw. Maintenance, ICSM*, pp. 70–79, 2013.
- [33] R. Chanajitt, W. Viriyasitavat, and K. K. R. Choo, "Forensic analysis and security assessment of Android m-banking apps," *Aust. J. Forensic Sci.*, vol. 50, no. 1, pp. 3–19, 2018.
- [34] J. Due, "MARA - A Mobile Application Reverse Engineering And Analysis Framework." *hacking.reviews*, 2017.
- [35] D. G. N. Benitez-Mejia, G. Sanchez-Perez, and L. K. Toscano-Medina, "Android applications and security breach," 2016 3rd International Conference on Digital Information Processing, Data Mining, and Wireless Communications, DIPDMWC 2016, pp. 164–169, 2016.
- [36] T. Chothia, F. D. Garcia, C. Heppel, and C. M. M. Stone, "Why banker bob (Still) Can't Get TLS right: A security analysis of TLS in leading UK banking apps," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 10322 LNCS, pp. 579–597, 2017.
- [37] E. Filiol and P. Irolla, "(In)Security of Mobile Banking and of Other Mobile Apps," *Black Hat Asia*, pp. 1–22, 2015.
- [38] X. Zheng, L. Pan, and E. Yilmaz, "Security analysis of modern mission critical android mobile applications," *ACM Int. Conf. Proceeding Ser.*, no. October, 2017.
- [39] B. Reaves, N. Scaife, A. Bates, P. Traynor, and K. R. B. Butler, "Mo(bile) Money, Mo(bile) Problems: Analysis of branchless banking applications in the developing world," *Proceedings of the 24th USENIX Security Symposium*, pp. 17–32, 2015.
- [40] D. Bassolé, G. Koala, Y. Traoré, and O. Sié, "Vulnerability Analysis in Mobile Banking and Payment Applications on Android in African Countries," *Lecture Notes of the Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering, LNICST*, vol. 321 LNICST, pp. 164–175, 2020.
- [41] S. A. Chaudhry, M. S. Farash, H. Naqvi, and M. Sher, "A secure and efficient authenticated encryption for electronic payment systems using elliptic curve cryptography," *Electron. Commer. Res.*, vol. 16, no. 1, pp. 113–139, 2016.
- [42] M. G. Manoti, "Enhancing Security of Mobile Banking and Payments in Kenya," no. November, 2016.
- [43] G2, "NNT Vulnerability Tracker," 2020. [Online]. Available: <https://www.g2.com/products/nnt-vulnerability-tracker/reviews>. [Accessed: 17-Nov-2020].
- [44] M. A. Hassan and Z. Shukur, "Device Identity-Based User Authentication on Electronic Payment System for Secure E-Wallet Apps," *Electronics.*, vol. 11, no. 1, pp. 1–29, 2022.
- [45] L. Verderame, D. Caputo, A. Romdhana, and A. Merlo, "On the (Un)Reliability of Privacy Policies in Android Apps," *Proc. Int. Jt. Conf. Neural Networks*, 2020.
- [46] S. S. Majeti, B. Janet, and N. P. Dhavale, "Analysis of Inappropriate Usage of Cryptographic Primitives in Indian Mobile Financial Applications," *Lecture Notes on Data Engineering and Communications Technologies*, vol. 56, pp. 211–220, 2021.

Demand Forecasting Model using Deep Learning Methods for Supply Chain Management 4.0

Loubna Terrada, Mohamed El Khaili, Hassan Ouajji
IESI Laboratory, ENSET Mohammedia
Hassan II University of Casablanca
Mohammedia, Morocco

Abstract—In the context of Supply Chain Management 4.0, costumers' demand forecasting has a crucial role within an industry in order to maintain the balance between the demand and supply, thus improve the decision making. Throughout the Supply Chain (SC), a large amount of data is generated. Artificial Intelligence (AI) can consume this data in order to allow each actor in the SC to gain in performance but also to better know and understand the customer. This study is carried out in order to improve the performance of the demand forecasting system of the SC based on Deep Learning methods, including Auto-Regressive Integrated Moving Average (ARIMA) and Long Short-Term Memory (LSTM) using historical transaction record of a company. The experimental results enable to select the most efficient method that could provide better accuracy than the tested methods.

Keywords—Supply chain management 4.0; demand forecasting; decision making; artificial intelligence; deep learning; Auto-Regressive Integrated Moving Average (ARIMA); Long Short-Term Memory (LSTM)

I. INTRODUCTION

Demand forecasting is one of the crucial challenges of demand planning in the Supply Chain Management that uses historical demands or sales data to predict future costumers' demands and support decision making [1]. It aims to enhance the logistics performance by optimizing stocks' value, minimizing costs and increasing sales to warrant the costumers' satisfaction [2].

The Smart Supply Chain or Supply Chain Management 4.0 (SCM 4.0) is a new paradigm introduced to solve this complexity by the integration of Artificial Intelligence (AI) [3]. AI has been implemented in several stages along the Supply Chain (SC) and showed a huge potential to impact the upstream or the downstream of the SC, to find smart solution to complex problems and deploy the massive amount of data generated at each stage [4].

Artificial Intelligence (AI) and Machine learning (ML) techniques are widely used in several fields. Deep learning (DL) is one of its most used methods, which is deployed mostly for time series problems, for instance: Urban Traffic Control [5], Production and Energy [6] [7], tasks scheduling and e-commerce [8], Smart Cities [9], Healthcare [10], Trading and Stock Price Predictions [11] [12].

The aim of our research emphasizes the role of AI in the Supply Chain by developing a smart demand forecasting

system based on Deep Learning methods to make the SC smart, collaborative and communicative [13].

Many studies have applied multiple types of Neural-Network models, such as Artificial Neural Networks (ANN) and Recurrent Neural Networks (RNN), in different areas, such as inventory management and distribution. However, few studies address the issue of demand forecasting in this context. At the same time, the current trend in methods for calculating forecasts, in many fields of activity, is towards Machine Learning approaches. They demonstrate that these models can dominate statistical methods such as linear regression and Auto-Regressive Integrated Moving Average (ARIMA). As statistical methods are theoretically linear models, they do not cope well with uncertainty and fluctuations in demand. However, few researchers use Long Short-Term Memory (LSTM) in demand forecasting [14] knowing that LSTM has shown relevant forecasting result in many fields.

This paper is structured as follows: Section II "Related work" incorporates demand forecasting of logistics and Supply Chain Management 4.0 in general, we also give an overview on the AI deployed in SCM field. In Section III, we describe the methodology adopted in our forecasting system, more precisely, the proposed models related to LSTM and ARIMA. In this part, we explain all the process and steps followed to define the proposed model. Section IV details an experiment with the proposed method used for comparison between LSTM and other Forecast time series method, such as, ARIMA. Then, we analyze and compare the results obtained in order to validate the most efficient DL method in terms of accuracy and performance. Finally, Section V, conclude the article with a brief overview on our future research perspectives.

II. RELATED WORK

A. Supply Chain Management 4.0: Outlooks

The SCM's principle is to ensure full cooperation and coordination between all the stockholders by developing consistent interactions, collaboration and coordination to achieve overall performance until the final customer [15]. The concept of Supply Chain emerged to warrant products' availability to customers by creating values throughout the whole process. Nevertheless, the SC has always been dealing with several issues such as uncertainty in forecasting and planning, as each stage in the SC requires a high-level accuracy in order to control inventory changes and to avoid

over-stocks and stock-outs. In the literature, this phenomenon is called "Bullwhip effect" [2]. The classic forecasting methods, implemented in many industries, have reached their limits. They are not able to deal with fluctuations in demand or take into account of the complexity of increasingly connected SC networks. Consequently, companies should migrate to intelligent systems and move towards a Smart Supply Chain Management.

According to the literature SCM 4.0 is defined as the interaction between advanced digital technologies and SCM, such as; Big Data, Artificial Intelligence, Cloud Computing, and Blockchain. Researchers have addressed the sustainability challenge through SCM4.0 and showed the impact of digital transformation technologies on SC sustainability to warrant better customers' experiences. Researchers proposed also a SCM4.0 Framework to define the main key topics of this field and the components for its development [3] as shown in Fig. 1.

B. Supply Chain Management 4.0: Challenges

SCM is facing many challenges, such as, demand forecast uncertainty. The latter has a significant effect on planning systems, uncertain demand leads to regular updating of system parameters and regular changing of targets [16]. It can be presented as the range in which the actual demand will continue. The forecast will very rarely be accurate. However, it gives a good idea of the actual demand. Thus, by adding this confidence interval to the forecast, we obtain more accurate information on the likely value of demands. This confidence interval will then be modelled on the product history in order to match reality as closely as possible. Moreover, this error is due to two effects: the quantity or the delay inconsistency. However, both cases have the same consequences: excess stock or shortage. In addition, these consequences can become even more serious when they are amplified by the "Bullwhip effect" [17]. This effect, illustrated in Fig. 2, describes the phenomenon whereby a small variation in demand at customer level will tend to increase throughout the Supply Chain, consequently, their operation inefficiency.

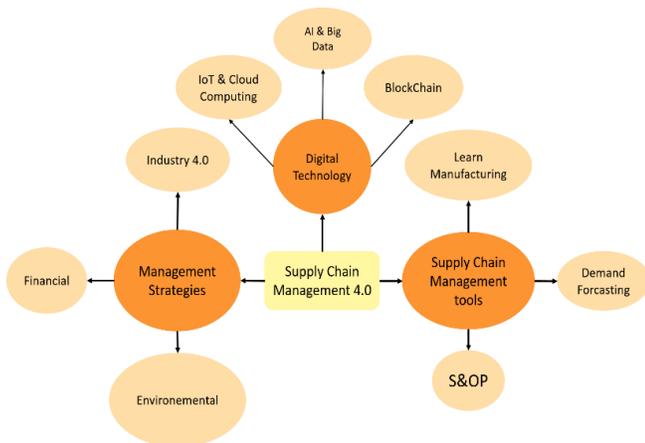


Fig. 1. Supply Chain Management 4.0 Framework.

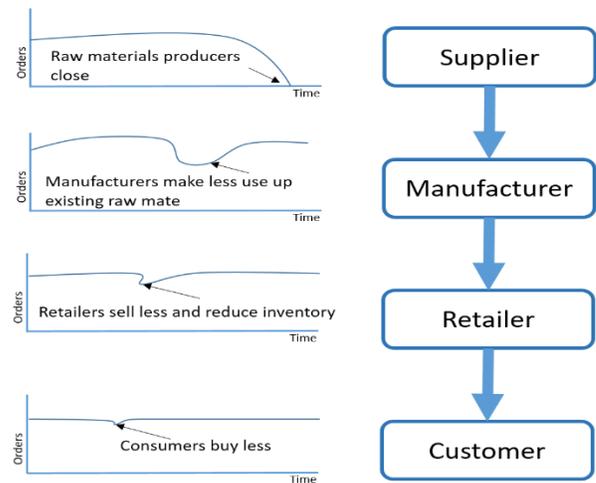


Fig. 2. Demand Distortion in the Supply Chain.

Thus, an intelligent demand forecasting system based on AI is the only solution to deal with the demand variation and then minimize the Bullwhip effect. This phenomenon is due to the Lead-time between the order and the delivery of goods, and Forecast changes that might occur. Each order has to adapt not only to fluctuations in demand during the current period, but also to changes in the level of predicted demand in the lead-time [18].

There are various AI methods used in demand prediction techniques in the literature. Deep Learning shows better performance and results in terms of forecasting comparing to other methods. Although the forecasting based only on the historical data of manufacturing demand is achievable, the accuracy of the prediction results is considerably lower than when taking into account multiple factors [19]. Researchers set up strategies of the exponential forecasting framework for sales forecasting in order to optimize manufacturing planning and inventory [20]. Therefore, our aim is to forecast future demands based on historical data coming from past manufacturer orders and retailers' demands based on customers' requirements, just enough time in advance to compensate for manufacturer Lead-times.

In the literature, researchers used several ML and DL methods to develop the accuracy of demand forecasting. Abbasimehr et al. used multi-layers LSTM method to compare with other time series forecasting methods, such as K-nearest neighbors (KNN), exponential smoothing (ETS), Auto-Regressive Integrated Moving Average (ARIMA), Support Vector Machines (SVM), Recurrent Neural Network (RNN), Artificial Neural Network (ANN) and Long-Short Term Memory (LSTM). The results of this study shows that LSTM method is more efficient compared to the tested methods with regards to performance measures [21]. Zixin et al. used LSTM, and Grey Model (GM) models in order to predict the future values based on historical data and the total industrial value. The Statistical Yearbook of GD was selected to represent the demand of the manufacturing industry. Other indicators are from this statistical yearbook from 2005 to 2020. The experiment shows that LSTM has excellent results in the previous comparative experiments. The GM model is the classic model in the field of Auto-regression. Nevertheless,

it does not give accurate results, which is significantly lower than the forecasting results given by LSTM considering multiple factors [19]. A study conducted by Raizada et al. is based on a comparative analysis of several Supervised Machine Learning algorithms, such as, Support Vector Machine (SVM), Random Forest Regression, K-NN Algorithm and Extra Tree Regression to build a forecasting model for future sales of 45 retail outlets of Walmart store in India. The study shows that that Extra Tree Regression Technique is the most efficient model to predict the sales for the selected dataset; however the predictions obtained from the algorithm may vary based on the variance in training data [22]. Jiaying et al. compared the performance of classical forecasting models and the latest developing forecasting technologies for perishable products and non-perishable items of a large grocery retailer. The Authors made a comparison in terms of performance and accuracy between many algorithms, such as: ARIMA, SVM, RNN and LSTM. The study shows that SVM, RNN and LSTM have a high predictive performance to for perishable products, whereas ARIMA is outstanding in the runtime and LSTM is the most efficient method to deal with non-perishable items due to its advanced prediction performance [23].

There are several ways to apply demand forecasting. In general, the forecasts fluctuations depends on the model used. Using multiple forecasting models could also highlight differences in forecasts. These differences may indicate the need for more research or better data input.

According to the findings, we assume that LSTM and ARIMA are the most efficient Deep Learning methods to warrant a high accuracy level for demand forecasting in the Supply Chain and to deal with its fluctuation to defeat the bullwhip effect.

III. PROPOSED METHODOLOGY

A. Conceptual Framework Description

This paper points out that existing research can provide a rich literature for demand forecasting models in manufacturing, to which we can refer for the selection of the prediction model in this research work.

Furthermore, although RNN algorithms are easier to fit complex non-linear relationships, their accuracy is inherently affected by many factors, such as: vanishing Gradient problem. Therefore, Long Short-Term Memory (LSTM) networks are suitable for demand forecasting in manufacturing, and they are the best to deal with vanishing Gradient problem. To verify the accuracy of the LSTM network, various prediction models are used as comparison models [24]. The selected methods will be used according to the SCM Business Process Model Notation (BPMN) [25] illustrated in Fig. 3. To build our generic model we referred to Supply Chain Operations Reference (SCOR) model and we used BPMN as a tool for modelling. We consider a SC BPM where each process is modelled by a separate pool and process chain as follows: Supplier, Manufacturer, Retailer and Customer. The interactions between the four Agents is managed and submitted to several flows and probabilities. The Agents cooperation, collaboration and coordination are

Crucial in the making-decision process, as efforts will only succeed if internal coordination, information exchange and material flow are effective.

B. Methods and Materials

In this sub-section, we give a brief description forecasting models used in our study.

1) *ARIMA*: Auto-Regressive Moving Average (ARIMA) models include three main procedures: Auto-Regression, Integration and Moving Average [26]. ARIMA can perform modeling of several kinds of time series. However, ARIMA’s limitation is that it assumes that the given time series is linear [21]. This model will be used in our study to be compared with LSTM to evaluate the efficiency of our proposed forecasting system.

In this process, the parameters of the Auto-Regressive Moving Average (ARIMA) model shown in Equation (1) are determined as (p) and (q) , respectively. An ARIMA model is defined as (p, d, q) .

- p : number of Auto-Regressive terms.
- d : degree of differencing.
- q : number of lagged forecast errors in the prediction equation (MA).

$$Y_t = \alpha_1 w_{t-1} + \alpha_2 w_{t-2} + \dots + \alpha_p w_{t-p} + \epsilon_t - \theta_1 \epsilon_{t-1} - \theta_2 \epsilon_{t-2} - \dots - \theta_q \epsilon_{t-q} \quad (1)$$

2) *LSTM*: Long-Short Time Memory (LSTM) method is gradient-based learning algorithm [24]. As illustrated in Fig. 4 [27]. The memory cell’s content is modeled by “Forget Gate”, “Input Gate” and “Output Gate”.

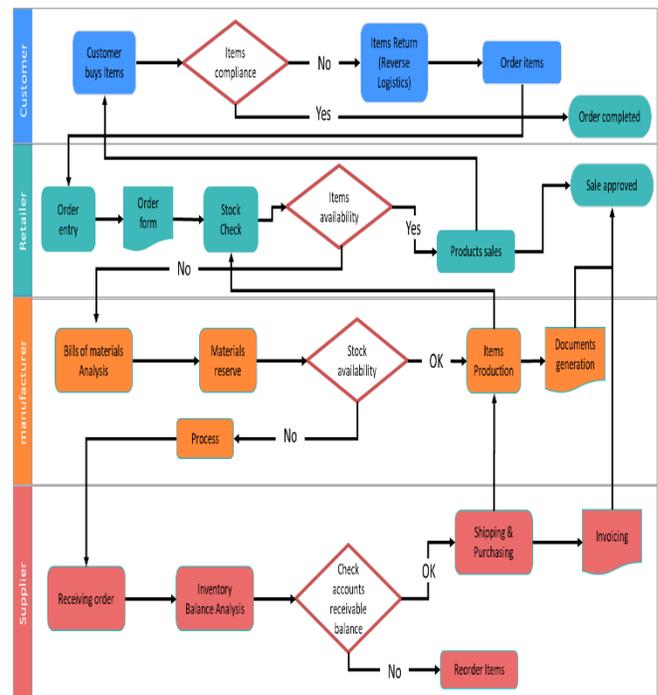


Fig. 3. SCM Business Model Process.

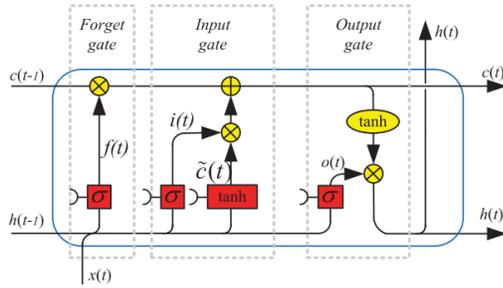


Fig. 4. LSTM Architecture.

LSTM Model notations are as follows:

- $x(t)$: represents the input value.
- $h(t-1)$: represents the output value at time t-1.
- $h(t)$: represents the output value at time t.
- $c(t-1)$: represents the cell state (memory) at time t-1.
- $c(t)$: represents the cell state (memory) at time t.
- $i(t)$: represents the Input Gate.
- $f(t)$: represents the Forget Gate.
- $o(t)$: represents the Output Gate.
- $W1$: represents $\{W_i, W_f, W_c, W_o\}$ weight matrixes.
- $W2$: represents $\{W_{ih}, W_{fh}, W_{ch}, W_{oh}\}$ the recurrent weights.
- b : represents $\{b_i, b_f, b_c, b_o\}$ biases for the gates.
- σ : represents the Sigmoid function.

Based on $x(t)$ and $h(t-1)$ the Forget Gate $f(t)$ can decide what information will be preserved in the cell state using as inputs using Sigmoid activation σ . The Input Gate $i(t)$ uses the input values $x(t)$ and $h(t-1)$ to compute the value of the cell $c(t)$. While the Output Gate $o(t)$ determines the output value $h(t)$ using $h(t-1)$, $x(t)$ and Sigmoid activation σ , whereas, \tanh activation function is used to compute the value of $c(t-1)$ and multiplies it to get $h(t-1)$.

The LSTM cell can be mathematically modelled as follows:

$$i(t) = \sigma(W_{ih} h_{t-1} + W_i x(t) + b_i) \quad (2)$$

$$f(t) = \sigma(W_{fh} h_{t-1} + W_f x(t) + b_f) \quad (3)$$

$$c(t) = f_i \cdot c_{t-1} + i_t \tanh(W_c x(t) + W_{ch} h(t-1) + b_c) \quad (4)$$

$$o(t) = \sigma(W_{oh} h_{t-1} + W_o x(t) + b_o) \quad (5)$$

$$h(t) = o(t) \cdot \tanh(c(t)) \quad (6)$$

Such that \tanh and σ are activation functions. By the evoked iteration and Compute the LSTM output using Equation (2)–(6), with the $x(t)$ Input, the model can compute the future value of the Output $o(t)$.

C. Research Framework

The aim of our research is to select an accurate model for demand forecasting based on the selected Dataset. We deploy the evoked DL methods to select the best time series forecasting model to deal with demand forecasting issues and uncertainty. The proposed methodology is summarized in Fig. 5. The flowchart emphasizes the main steps of our method from the Data collection up to the Output predicted value.

The five main steps of our methodology are detailed in the following sub-sections.

1) *Data collection and pre-processing*: In this research, we use a dataset from Kaggle's competition (<https://www.kaggle.com/code/devswaroop/forecastproductsdemand/data>) which is suits our proposed generic model. Data collection is a crucial step because the quality and volume of data. It's a success factor of the predictive system. In this case, the data used in this study will be the risk factors of Supply Chain components. This will yield us a table of different Features. The more data we collect, the more accuracy we can get while avoiding over-fitting effect [28].

The inputs selection should be in concordance with the system's objective and problem that we need to solve. In our study, we aim to implement a LSTM based forecasting system to predict demand quantities of a specific product based on past values and compare them with the results obtained through ARIMA. The dataset used in our experiment contains demand quantities of several products from 2011 to 2017. Consequently, the neural network's output is the estimated demand and the previous demand quantities with the month' classification implemented as inputs in the input layer. We load our data into a suitable place for pre-processing before using it on our DL models (see Table I).

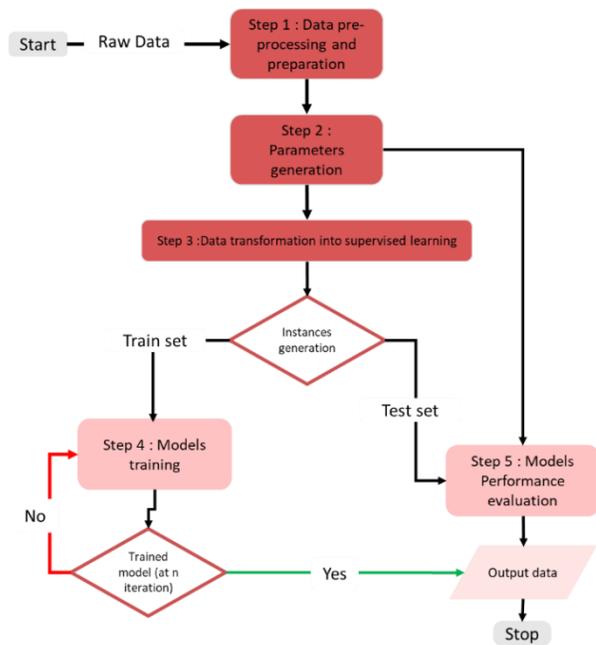


Fig. 5. Our Proposed Methodology Flowchart.

TABLE I. DATASET SAMPLE DEPLOYED DURING EXPERIMENTATIONS

Quantity per Month	Inputs				Warehouse	Output
	Demand				Average	Demand
Product Category	X1	X2	X3	Xm	Average	Y
Product1	X ¹ ₁	X ¹ ₂	X ¹ ₃	X ¹ _m	Whse_J	Y ₁
Product2	X ² ₁	X ² ₂	X ² ₃	X ² _m		Y ₂
Product3	X ³ ₁	X ³ ₂	X ³ ₃	X ³ _m	Whse_S	Y ₃
Product4	X ⁴ ₁	X ⁴ ₂	X ⁴ ₃	X ⁴ _m		Y ₄
Product5	X ⁵ ₁	X ⁵ ₂	X ⁵ ₃	X ⁵ _m	Whse_C	Y ₅
Product6	X ⁶ ₁	X ⁶ ₂	X ⁶ ₃	X ⁶ _m	Whse_A	Y ₆
Product _n	X ⁿ ₁	X ⁿ ₂	X ⁿ ₃	X ⁿ _m		Y _m

We split the dataset into two subsets: training set (80%) and Test set (20%). Since the dataset was monthly demand data, 1-month-ahead (one-step-ahead), forecasting was performed. Fig. 6 illustrates the time-series evolution of product demand per month.

To prepare the dataset for the training model, we followed the following steps for missing value processing and convert the original data days into months:

- Remove the missing value: remove lines of missing values from the analysis sample.
- Average interpolation: observe the average instead of the missing values.
- High frequency data: refers to time-series data collected at an extremely fine scale. It could be accurately collected at an efficient rate for analysis.

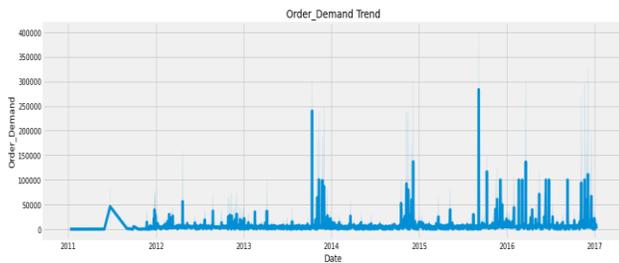


Fig. 6. Order Demand Evolution from 2011 to 2017.

Fig. 7 illustrates the volume of demand for each product category after data cleaning and removing the outliers.

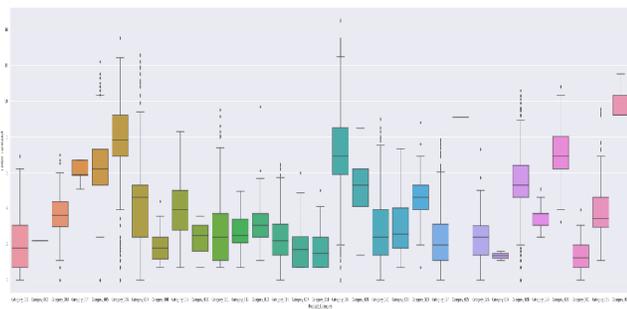


Fig. 7. Demand Volume per Product Category.

2) *Evaluation criteria:* To measure the performance of the proposed method, we used Mean Squared Error (MSE), Root Mean Square Error (RMSE), Mean Absolute Error (MAE) and Mean Absolute Percentage Error (MAPE) The LSTM and ARIMA were implemented and trained using Scikit-learn package and Keras in Python. For evaluation, we use MSE, RMSE, MAE and MAPE models defined as follows:

$$MSE = \frac{1}{n} \sum_{t=1}^n \frac{1}{n} (y_t - \hat{y}_t)^2 \tag{7}$$

$$RMSE = \sqrt{\frac{1}{n} \sum_{t=1}^n \frac{1}{n} (y_t - \hat{y}_t)^2} \tag{8}$$

$$MAE = \left(\sum_{t=1}^n \left| \frac{y_t - \hat{y}_t}{y_t} \right| \right) \tag{9}$$

$$MAPE = \left(\sum_{t=1}^n \left| \frac{y_t - \hat{y}_t}{y_t} \right| \right) \frac{100}{n} \tag{10}$$

In Equations (7)-(10), y_t indicates the real value, whereas \hat{y}_t is the predicted value, and n is the number of forecast periods. The model with the lowest standard value obtained using the above metrics should be selected as the most suitable and efficient model for the dataset.

IV. RESULT AND DISCUSSION

In this section, we provide results of our experiment based on the selected dataset.

A. ARIMA Model Findings

ARIMA Model (p, d, q) is implemented using $p [0, 1, 2]$, $d [0, 1, 2]$, $q [0, 1, 2]$ values for the demand data forecasting, and Correlogram test of each of these models were performed apart. Table II indicates the results of several parameters and performance metrics comparison between models. According to the obtained results, the model with a lower error rate is selected as the model with a higher performance level. In this regard, ARIMA (2,2,2) is the model with the lowest MAPE value, thus, it could provide the most accurate forecast among ARIMA models that we performed.

Fig. 8 illustrate the 12 Months forecast data obtained using the model ARIMA (2,2,2) with the lowest error rate value versus actual test data.

TABLE II. ARIMA MODELS CORRELOGRAM RESULTS

Comparison of ARIMA Models				
ARIMA	MSE	RMSE	MAE	MAPE
(1,1,1)	3.70	608840.38	352746.12	2.04
(0,1,0)	4.55	675233.37	420006.19	2.27
(1,0,0)	4.25	652650.05	352531.14	2.28
(1,1,0)	4.53	673113.87	390188.60	2.31
(1,0,1)	4.19	648035.65	356377.35	2.27
(0,0,2)	4.19	647642.11	350322.06	2.24
(2,2,2)	1.67	408946.67	317530.90	0.75

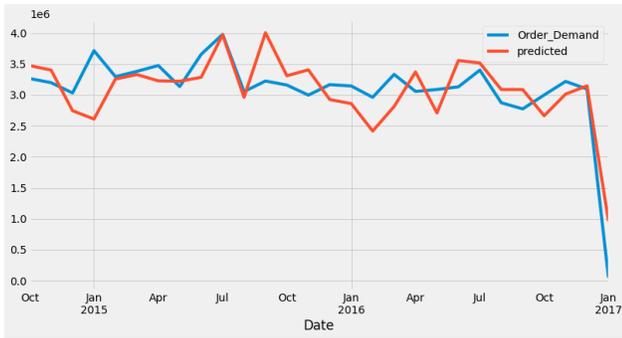


Fig. 8. ARIMA (2, 2, 2) Model and Dataset Comparison Chart.

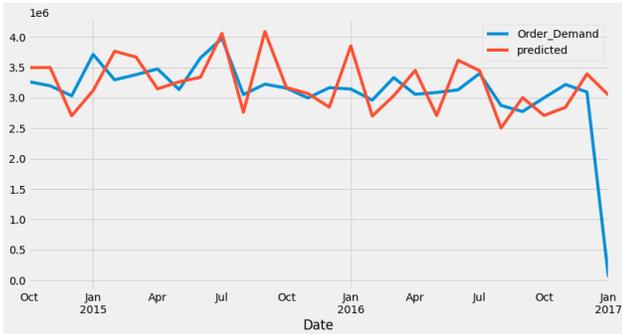


Fig. 9. ARIMA (0, 1, 0) Model and Dataset Comparison Chart.

Analysis of Fig. 8 and 9 shows that monthly demands values obtained from real data and estimated studies have veering structure and the deviation is not excessive. The efficiency of the model could be seen more clearly in the graph, than the similarity of breakpoint directions and the approximation of the data. The model used here produces values that are very close to the real data with an error of 0.75 MAPE. This situation suggests that the model used in this experiment was compliant.

B. LSTM Model Findings

In the second experiment, LSTM was trained with the dataset, using Python with KERAS. We run the LSTM model on the monthly demands of the products listed between 2011 and 2017, as done in the ARIMA model. We tried different epoch numbers in the training process, thus, we examined the error values results. The error rates obtained from epoch numbers according to the training combinations performed are indicated in Fig. 10.

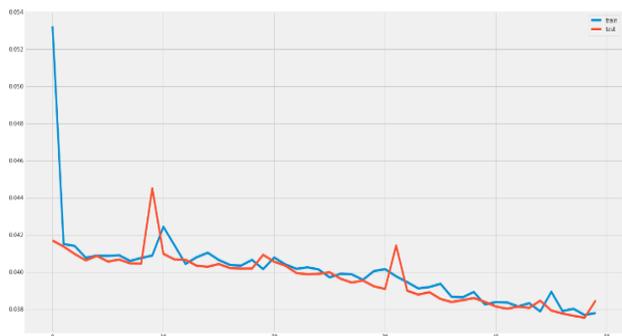


Fig. 10. Train Plot and Test Loss during LSTM Model Training.

According to the obtained results, LSTM model has managed to provide reliable results with the used data. It produced a lower error value compared to the ARIMA model with Epoch 500 with MSE and MAPE.

C. ARIMA versus LSTM

The time series data used in this experiment is monthly data. The Demand forecasting was performed using different models: ARIMA and LSTM.

Our aim is to train the models to select the best that can provide better accuracy for the used dataset and to compare the models' performance.

In Table III, we indicate the MSE values of ARIMA and LSTM calculated as 1.67 and 1.12 respectively. Considering the MSE and MAPE values, we assume that LSTM is the model suitable to the used dataset and could provide better results in terms of products demand prediction.

In this study, a forecasting was carried out for monthly housing data demand with the selected dataset using the DL methods evoked previously. The data in question was not only computed by being processed in the program just once, but the model was trained multiple times until realistic and reliable values were obtained. According to Table III, the error values that depict the performance metrics, for each method are considerably low. The forecasting accuracy in demand for retailers and manufacturer, with regards to a more balanced supply and demand, will provide reliable information and visibility on the future demand, thus help the Supply Chain stockholders in the making-decision process. The aim is to transform the traditional SCM into Smart Supply Chain Management. In addition, the method and dataset used in our experiment is matching the generic model that we proposed and might be applied by any company despite its size and environment anywhere in the world, because it does not consider a specific situation in the economic environment where the forecasting is made. In this respect, we proposed a generic model which is "oriented-Agent" [29] and also "oriented-process" based on SCOR and using BPMN as a modelling tool. Many different methods could be deployed for forecasting and it may be possible to provide different results from each method [30]. In this purpose, we used two methods in our study and the results obtained from each method were compared in terms of their proximity to the real values. According to the performance metrics of the forecasting models, we deduce that LSTM is more efficient and could provide better results. The demand forecasting is a crucial step in the upstream Supply Chain; it aims to control the Bullwhip effect and also to enhance the Key Performance throughout the Supply Chain from the supplier up to the final customer [31].

TABLE III. PERFORMANCE METRICS COMPARISON BETWEEN LSTM AND ARIMA

Applied method	MSE	RMSE	MAE	MAPE
Auto-Regressive Moving Average	1.67	408946.67	317530.90	0.75
Long-Short Time Memory	1.12	3421527.86	3375479.27	0.65

V. CONCLUSION

In this paper, we focused on one of the main Supply Chain issues related to decision-making, fluctuation and uncertainty of information flow and demand; we have focused on the "Bullwhip effect" phenomenon in order to propose advanced solutions to reduce it. Accurate forecasts are mandatory to improve the performance key indicators of the Supply Chain. Providing a wider range of data, information sharing and collaborative forecasting are crucial in order to enable the supply chain to increase profitability and minimize waste or delays. Similarly, negative data could also lead to downward changes in the statistical forecasts, which could lead to downward changes in forecasting.

In our case study, we deployed two Deep Learning models ARIMA and LSTM, to build our demand forecasting system and to deal with regression problems. We used a dataset collected from Kaggle whose Features correspond best to the generic model of Supply Chain that we proposed which is "Agent-oriented" and "process-oriented". Our experiment and training have shown that multi-layer LSTM gives estimated values closer to reality than those obtained by ARIMA. In particular, LSTM models perform better because they allow better persistence of information compared to classical RNNs and ARIMA, due to the information transmission over time by the hidden state called the "cell state". The aim of our approach is to maintain the balance between the supply and demand in the Supply Chain, thus, incorporate intelligent predicting system using AI, which is a crucial component of Supply Chain Management 4.0.

As perspective of this study, we will propose a Hybrid forecasting model based on ARIMA and LSTM to enhance the performance of our predicting system and improve the accuracy of the results.

REFERENCES

- [1] J. Feizabadi, "Machine learning demand forecasting and supply chain performance," *International Journal of Logistics Research and Applications*, vol. 25, no. 2, pp. 119–142, Feb. 2022, doi: 10.1080/13675567.2020.1803246.
- [2] E. Hofmann and E. Rutschmann, "Big data analytics and demand forecasting in supply chains: a conceptual analysis," *IJLM*, vol. 29, no. 2, pp. 739–766, May 2018, doi: 10.1108/IJLM-04-2017-0088.
- [3] K. Zekhnini, A. Cherrafi, I. Bouhaddou, Y. Benghabrit, and J. A. Garza-Reyes, "Supply chain management 4.0: a literature review and research framework," *BIJ*, vol. 28, no. 2, pp. 465–501, Sep. 2020, doi: 10.1108/BIJ-04-2020-0156.
- [4] H. Min, "Artificial intelligence in supply chain management: theory and applications," *International Journal of Logistics Research and Applications*, vol. 13, no. 1, pp. 13–39, Feb. 2010, doi: 10.1080/13675560902736537.
- [5] M. El Khaili, L. Terrada, H. Ouajji, and A. Daaif, "Towards a Green Supply Chain Based on Smart Urban Traffic Using Deep Learning Approach," *Stat., optim. inf. comput.*, vol. 10, no. 1, pp. 25–44, Feb. 2022, doi: 10.19139/soic-2310-5070-1203.
- [6] F. Abdullayeva and Y. Imamverdiyev, "Development of Oil Production Forecasting Method based on Deep Learning," *Stat., optim. inf. comput.*, vol. 7, no. 4, pp. 826–839, Dec. 2019, doi: 10.19139/soic-2310-5070-651.
- [7] G. H. Alraddadi and M. T. B. Othman, "Development of an Efficient Electricity Consumption Prediction Model using Machine Learning Techniques," *IJACSA*, vol. 13, no. 1, 2022, doi: 10.14569/IJACSA.2022.0130147.
- [8] Y. Issaoui, A. Khiat, A. Bahnasse, and H. Ouajji, "An Advanced LSTM Model for Optimal Scheduling in Smart Logistic Environment: E-Commerce Case," *IEEE Access*, vol. 9, pp. 126337–126356, 2021, doi: 10.1109/ACCESS.2021.3111306.
- [9] M. El Khaili, L. Terrada, A. Daaif, and H. Ouajji, "Smart Urban Traffic Management for an Efficient Smart City," in *Explainable Artificial Intelligence for Smart Cities*, 1st ed., Boca Raton: CRC Press, 2021, pp. 197–222. doi: 10.1201/9781003172772-11.
- [10] O. Terrada, B. Cherradi, A. Raihani, and O. Bouattane, "A novel medical diagnosis support system for predicting patients with atherosclerosis diseases," *Informatics in Medicine Unlocked*, vol. 21, p. 100483, 2020, doi: 10.1016/j.imu.2020.100483.
- [11] Y. Touzani and K. Douzi, "An LSTM and GRU based trading strategy adapted to the Moroccan market," *J Big Data*, vol. 8, no. 1, p. 126, Dec. 2021, doi: 10.1186/s40537-021-00512-z.
- [12] M.-L. Thormann, J. Farchmin, C. Weisser, R.-M. Kruse, B. Säfken, and A. Silbersdorff, "Stock Price Predictions with LSTM Neural Networks and Twitter Sentiment," *Stat., optim. inf. comput.*, vol. 9, no. 2, pp. 268–287, May 2021, doi: 10.19139/soic-2310-5070-1202.
- [13] L. Terrada, J. Bakkoury, M. El Khaili, and A. Khiat, "Collaborative and Communicative Logistics Flows Management Using the Internet of Things," in *Lecture Notes in Real-Time Intelligent Systems*, vol. 756, J. Mizera-Pietraszko, P. Pichappan, and L. Mohamed, Eds. Cham: Springer International Publishing, 2019, pp. 216–224. doi: 10.1007/978-3-319-91337-7_21.
- [14] T. Weng, W. Liu, and J. Xiao, "Supply chain sales forecasting based on lightGBM and LSTM combination model," *IMDS*, vol. 120, no. 2, pp. 265–279, Sep. 2019, doi: 10.1108/IMDS-03-2019-0170.
- [15] L. Terrada, A. Alloubane, J. Bakkoury, and M. E. Khaili, "IoT contribution in Supply Chain Management for Enhancing Performance Indicators," in *2018 International Conference on Electronics, Control, Optimization and Computer Science (ICECOCS)*, Kenitra, Dec. 2018, pp. 1–5. doi: 10.1109/ICECOCS.2018.8610517.
- [16] M. C. Cooper, "Issues in Supply Chain Management," *Industrial Marketing Management*, vol. 29, no. 1, pp. 65–83, Jan. 2000, doi: 10.1016/S0019-8501(99)00113-3.
- [17] H. L. Lee, V. Padmanabhan, and S. Whang, "Information Distortion in a Supply Chain: The Bullwhip Effect," *Management Science*, vol. 43, no. 4, pp. 546–558, Apr. 1997, doi: 10.1287/mnsc.43.4.546.
- [18] K. Gilbert, "An ARIMA Supply Chain Model," *Management Science*, vol. 51, no. 2, pp. 305–310, Feb. 2005, doi: 10.1287/mnsc.1040.0308.
- [19] Z. Dou, Y. Sun, Y. Zhang, T. Wang, C. Wu, and S. Fan, "Regional Manufacturing Industry Demand Forecasting: A Deep Learning Approach," *Applied Sciences*, vol. 11, no. 13, p. 6199, Jul. 2021, doi: 10.3390/app11136199.
- [20] H. Younis, B. Sundarakani, and M. Alsharairi, "Applications of artificial intelligence and machine learning within supply chains: systematic review and future research directions," *JM2*, Aug. 2021, doi: 10.1108/JM2-12-2020-0322.
- [21] H. Abbasimehr, M. Shabani, and M. Yousefi, "An optimized model using LSTM network for demand forecasting," *Computers & Industrial Engineering*, vol. 143, p. 106435, May 2020, doi: 10.1016/j.cie.2020.106435.
- [22] S. Raizada and J. R. Saini, "Comparative Analysis of Supervised Machine Learning Techniques for Sales Forecasting," *IJACSA*, vol. 12, no. 11, 2021, doi: 10.14569/IJACSA.2021.0121112.
- [23] J. Wang, G. Q. Liu, and L. Liu, "A Selection of Advanced Technologies for Demand Forecasting in the Retail Industry," in *2019 IEEE 4th International Conference on Big Data Analytics (ICBDA)*, Suzhou, China, Mar. 2019, pp. 317–320. doi: 10.1109/ICBDA.2019.8713196.
- [24] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997, doi: 10.1162/neco.1997.9.8.1735.
- [25] K. Grzybowska and G. Kovács, "The modelling and design process of coordination mechanisms in the supply chain," *Journal of Applied Logic*, vol. 24, pp. 25–38, Nov. 2017, doi: 10.1016/j.jal.2016.11.011.
- [26] G. E. P. Box, G. M. Jenkins, G. C. Reinsel, and G. M. Ljung, *Time series analysis: forecasting and control*, Fifth edition. Hoboken, New Jersey: John Wiley & Sons, Inc, 2016.

- [27] Y. Yu, X. Si, C. Hu, and J. Zhang, "A Review of Recurrent Neural Networks: LSTM Cells and Network Architectures," *Neural Computation*, vol. 31, no. 7, pp. 1235–1270, Jul. 2019, doi: 10.1162/neco_a_01199.
- [28] T. Dietterich, "Overfitting and undercomputing in machine learning," *ACM Comput. Surv.*, vol. 27, no. 3, pp. 326–327, Sep. 1995, doi: 10.1145/212094.212114.
- [29] L. Terrada, M. E. Khaili, and H. Ouajji, "Multi-Agents System Implementation for Supply Chain Management Making-Decision," *Procedia Comput. Sci.*, vol. 177, pp. 624–630, 2020, doi: 10.1016/j.procs.2020.10.089.
- [30] A. R. S. Parmezan, V. M. A. Souza, and G. E. A. P. A. Batista, "Evaluation of statistical and machine learning models for time series prediction: Identifying the state-of-the-art and the best conditions for the use of each model," *Inf. Sci.*, vol. 484, pp. 302–337, May 2019, doi: 10.1016/j.ins.2019.01.076.
- [31] L. Terrada, A. Alloubane, J. Bakkoury, and M. E. Khaili, "IoT contribution in Supply Chain Management for Enhancing Performance Indicators," in *2018 International Conference on Electronics, Control, Optimization and Computer Science (ICECOCS)*, Kenitra, Dec. 2018, pp. 1–5. doi: 10.1109/ICECOCS.2018.8610517.

Improved Deep Learning Performance for Real-Time Traffic Sign Detection and Recognition Applicable to Intelligent Transportation Systems

Anass BARODI^{1*}, Abderrahim Bajit², Abdelkarim ZEMMOURI³, Mohammed Benbrahim⁴, Ahmed Tamtaoui⁵

Laboratory of Advanced Systems Engineering (ISA), National School of Applied Sciences
Ibn Tofail University, Kenitra, 14000, Morocco^{1,2,3,4}

National Institute of Posts and Telecommunications (INPT-Rabat)
SC Department, Mohammed V University, Rabat, 10000, Morocco⁵

Abstract—In this paper, we improve the performance of Deep Learning (DL) by creating a robust and efficient Convolutional Neural Network (CNN) model. This CNN model will be subjected to detecting and recognizing traffic signs in real-time. We apply several techniques; the first is pre-processing, which includes conversion of color space RGB, then equalization and normalization histogram of the image dataset according to Computer Vision (CV) tools. The second is devoted to Artificial Intelligence (AI), which needs the right choice of a neural layer such convolution layer, or dropout layer, with powerful optimizer as Adam and activation functions such as ReLU and SoftMax. Also, we use the technique of augmentation dataset which characterizes by the function of batch size for each epoch. The results obtained is very satisfactory, the prediction at the average is equal to 98%, which encourages this approach to the integration in Intelligent Transportation Systems (ITS) in the automotive sector.

Keywords—Deep learning; convolutional neural network; computer vision; artificial intelligence; traffic sign detection; traffic sign recognition; intelligent transportation systems

I. INTRODUCTION

The detection and recognition of traffic road signs are done in different ways, depending on the methodology or strategy followed by the researcher. In general, the detection and recognition methods can be summarized in three classes. The first method can be based on color segmentation (red, blue, yellow) [1]. In the second method, we can use the geometry of objects (Triangular, Square, Rectangle)[2]. Finally, methods that use artificial intelligence (AI), specifically DL of CNN architecture [3]. For road safety, we use ITS systems [4]. This system is devoted to detecting and recognizing all traffic road signs by identifying them from other objects that existed in environments (a passage, animals, cars, trucks, buildings.....) in real-time[5]. These systems are used in Advanced Driving Assistance Systems (ADAS) [6][7] and are based on a digital camera for perception road environment.

There is a standard technique for detecting and recognizing traffic road signs. For example, the scale-invariant feature transform (SIFT) [8][9], the local binary patterns (LBP) [10], and the histogram of oriented gradients (HOG) [11]. Also, we find advanced techniques to classify a different object, in which the feature vectors are extracted normally from the

training dataset, for example, the support vector machine SVM [12], VGG16 [13], and ImageNet [14]. In recent years, we have been using the CNN model for complex classification situations [15]. The CNN architectures are the best models; they have the same analysis vision as the human being. [16].

To guarantee a reliable and effective model in the decision, most of the research work in the field of AI often plays on the following parameters: optimizer [17], accuracy function [18], loss function [19], dataset [20], architectures [21].

In this paper, we play with several parameters to obtain a robust and efficient model for traffic sign detection and recognition. The first thing we will examine is the effect of normalizing and equalizing the images in the traffic sign dataset on model training. So according to the result of the first step, the second step is choosing an optimal fitting function (Simple, Generator) for deploying the best function between them. Finally, we will use the data augmentation technique by discussing the effect of batch size function during model training. All this is to ensure that the proposed ITS system detects and recognizes signs well in advance so the right decisions can be made as quickly as possible.

In our work, we will test our approach based on Computer Vision (CV) and Artificial Intelligence (AI), for the detection and recognition of the different traffic road signs in real-time. The approach results can be exploited by Intelligent Transport System (ITS) to assist the driver. The paper is organized as follows: Section 1 introduces the most techniques used for the detection and recognition of road signs. Section 2 is dedicated to related work, and then Section 3 presents a general view of the approach proposed. Section 4 is for methodology. Section 5 is devoted to experimentation and evaluation of the approach proposed. Section 6 with Section 7, is for the real-time implementation to recognize traffic road signs. The last section is devoted to the conclusion.

II. RELATED WORK

Most of the developed applications that have high accuracy in object detection and recognition are based on the RNN and CNN architecture [22]. Nevertheless, depending on the available data or the problem to be solved, one type of neural network may be more suitable and used than another for a different problem than the one it is used. Generally, a

Recurrent Neural Network (RNN) is used for text processing and speech recognition as illustrated in Table I. In this regard, convolution networks are applicable for object recognition in images and can specifically identify the shape of objects as illustrated in Table II. In this work, we will use the CNN architecture which is the most efficient neural network model concerning the available dataset.

A. The Constraints of the Traffic Sign Detection and Recognition Algorithms

Detection and recognition based on color segmentation are ranked as one of the fastest methods [41], applied for example to the recognition of road lanes, traffic signs, and vehicle license plates. Most algorithms use this technique to extract regions of interest, by setting specific filters to recognize apparent objects [42]. But this method can meet several problems such as weather conditions (snow, rain...), time of day (morning, night...) which has a great effect on the appearance (light reflection on the signs), or object distance (between the camera and road sign), lead to a false object detection recognition.

Some authors apply a more reliable method, it is the detection and recognition of the geometry of the road signs [43], that the detection is made on the basis of the objects contours in the image. To avoid any overlapping with the objects existing in the road environment by a structural analysis of the road signs [44].

TABLE I. MOST RNN ARCHITECTURE APPLICATIONS

Applications	RNN Architecture	Reference
Text processing	Efficient RNN Text Classification	J. Du [23] H.Chen [24] Z. Parcheta [25]
	Medical Text Classification Framework	X. Li [26] M. Ibrahim [27]
Speech recognition	Anticipation-RNN to Interactive Music Generation	F. Nielsen [28] D. Bisharad [29]
	Sentiment Analysis	A. Onan [30] J. Huan [31]

TABLE II. MOST CNN ARCHITECTURE APPLICATIONS

Applications	CNN Architecture	Reference
Image recognition	Traffic sign recognition systems	Á. Arcos-García [32] Á. Arcos-García [33]
	Lane Detection in Traffic Scene	J. Li [34] J. Kim [35] J. Tang [36]
Form recognition	CNN Design for Real-Time Traffic Sign Recognition	A. Shustanov [37] F. Shao [38]
	CNN Network for Real-life Traffic Sign Detection	T. Yang [39] Á. Arcos-García [40]

B. Deep Learning and Neural Network

The learning methods are among the techniques that use DL [45], this method has made a revolution in the industrial sector, especially in the embedded systems in the automotive sector [46]. This method is robust in object detection and recognition compared to the geometric and colorimetric methods, which are among the classic methods that suffer from many factors.

The creation of CNN models was based on neural networks. Many hidden layers of the neural network serve to produce CNN. These neuron layers are grouped into a tree category of layers, input layers, hidden layers, and output layers. Firstly, the feature vectors dataset is accepted from the input layer and has a bias neuron. Secondly, the liaison between input and output is hidden layers that use the neuron bias. Finally, the output of neural networks is not used for the bias neuron. The output from a single neuron is calculated according to the following equation (1).

$$f(x, \theta) = \phi(\sum_i(\theta_i \cdot x_i)) \quad (1)$$

- The input vector (x) represents the feature vector.
- The vector θ represents the weights.
- The ϕ is the transfer/activation function.

III. THE APPROACH DESCRIPTION

The approach that will be proposed to integrate it into an ITS system, is essentially based on the creation of a CNN architecture to guarantee road safety for both passengers and drivers of vehicles. Therefore, our approach is based on two processes, the detection process and the recognition of traffic signs as shown in Fig. 1. The detection process uses camera-based CV techniques to receive images to ensure that traffic signs are detected. When the signs are detected, the recognition process is activated using AI techniques. We will use a CNN architecture to extract the characteristics of the road signs. To achieve our objective, the deep training will be done on the German Traffic Sign Recognition Benchmark (GTSRB) dataset. We arrive at the end to identify each detected by the classes that belong to the prediction probability.

The strategy we will follow to have an efficient CNN architecture is summarized in the following points:

- Transformation techniques (equalization and normalization histogram).
- Creation of CNN architecture (convolution layers, and max-pooling layers)
- DL of the CNN architecture with simple fit function and generator fit function.
- Testing the performance of the CNN model in real-time detection and recognition of traffic road signs.

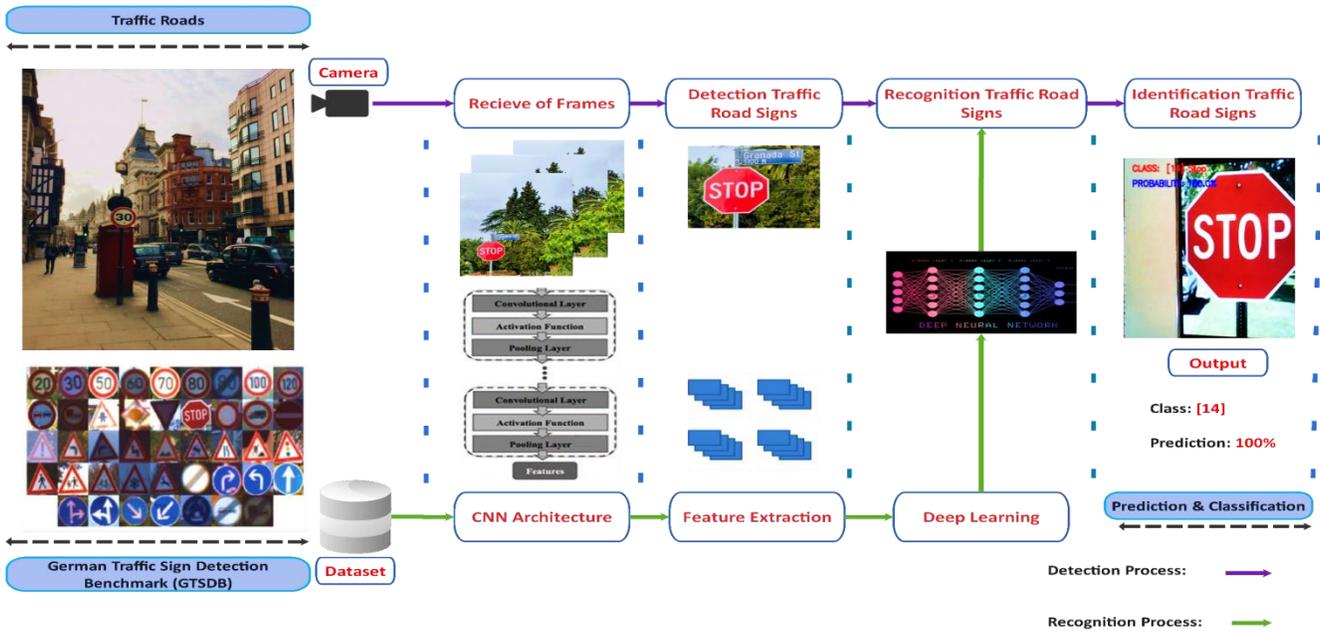


Fig. 1. General View of Approach Applicable for the ITS System.

IV. METHODOLOGY

A. Transformation Techniques for Dataset

a) *Visualization dataset:* For our implementation, we use a dataset of the German traffic sign Benchmark [47], composed of the tree part, training data, validation data, and testing data. The training set uses 80% of the data and the validation set uses 20%. The GTSRB is composed of 43 traffic road sign classes, 34799 images for training data, 4410 images for validation data, and 12630 images for testing data, as illustrated in Fig. 2.

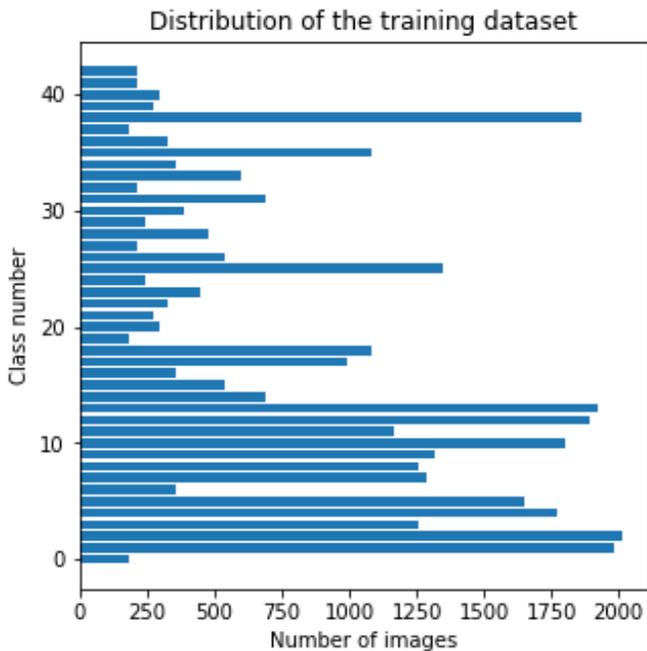


Fig. 2. Visualization of GTSRB Training Datasets.

b) *Normalization of a histogram:* Normalizing a histogram is a technique consisting of transforming the discrete distribution of intensities into a discrete distribution of probabilities [48]. To do this, we need to divide each value of the histogram by the number of pixels. In our case, the normalization is done by dividing all pixels in an image by 255.

c) *Equalization of a histogram:* Histogram equalization is an image processing method to adjust the contrast of an image, by modifying the intensity distribution of the histogram [49]. Equalization processing is based on the use of the cumulative probability function. This function is a cumulative sum of all the probabilities in its domain and is defined by equation (2).

$$cdf(x) = \sum_{k=-\infty}^x P(k) \quad (2)$$

The idea of this processing is to give the resulting image a linear cumulative distribution function.

B. Convolutional Neural Networks (CNNs)

Domain CV has been affected by AI mainly by CNNs. The neural network architecture was introduced by LeNet-5 [50]. The next step is the description of each layer type used in the CNN model.

a) *Convolution layers:* The first layer of analysis is the convolution, it allows us to detect the characteristics of each visual element: circles, lines, colors, edges ..., this work is done by internal filters in the layer. If the number of filters is very well brought, they have more features for better accuracy. The filters have a square shape that sweeps over the image from the right to the left. Then there is a very important parameter, which is the width and length of the filter that normally affects the number of features extracted from the images. The single

output matrix of the convolution layer is described in equation (3).

$$M_j = g(\sum_{i=1}^N Img_i * Ker_{i,j} + b_j) \quad \# \quad (3)$$

Img : Input matrix. Ker : Kernel matrix.

b_j : Bias. g : Non-linear activation.

Each set of kernel matrices represents a local feature extractor that extracts regional features from the input matrices. Optimizes neural network connection weights, and can be applied here to train the kernel matrices, biases as shared neuron connection weights.

b) Max pooling layers and dropout layers: Putting the Max-Pooling layers belong after every convolution layer. It serves for re-sizing a picture of 2D in a smaller dimension [51]. Most CNN frameworks implement dropout as a separate layer to avoid the production in DL the overfitting. Dropout layers function like a regular, densely connected CNN layer. The only difference is that the dropout layers will periodically drop some of their neurons during training.

c) Activation function: However, current deep neural networks mainly use the following activation functions, each function has a role to play in a neural network. For the output of the hidden layers, we use the ReLU (Rectified Linear Unit) function [52]. The ReLU function is calculated as follows in equation (4).

$$\phi(x) = \max(0, x) \quad \# \quad (4)$$

The ReLU activation function [53][54] was one of the key improvements in CNN applications, that make deep learning. Unfortunately, the ReLU function is not differentiable at the origin, which makes it hard to use with backpropagation training. ReLU for rectified the feature map, to find the final value positive and deleted the negative value.

The output of classification CNN: We implemented SoftMax. The SoftMax is calculated as follows in equation (5).

$$\phi_i(z) = \frac{e^{z_i}}{\sum_{j \in group} e^{z_j}} \quad \# \quad (5)$$

The SoftMax function is only useful with more than one output neuron. It guarantees that the sum of all output neurons is equal to 1.0. It is therefore very useful for classification, where it indicates the probability that each of the classes is the correct choice.

d) Optimization function: Adam optimizer is very effective [55]. Adam estimates the first mean and second variance moments to determine the weight corrections. Adam begins with an exponentially decaying average of past gradients (m) described in equation (6).

$$m_t = \beta_1 m_{t-1} + (1 - \beta_1) g_t \quad \# \quad (6)$$

g_t : the gradient at time t .

This average accomplishes a similar goal as a classic momentum update; however, its value is calculated automatically based on the current gradient (g). The update rule then calculates the second moment (v) in equation (7).

$$v_t = \beta_2 v_{t-1} + (1 - \beta_2) g_t^2 \quad \# \quad (7)$$

The values m_t and v_t are estimates of the first moment (the mean) and the second moment (the uncentered variance) of the gradients respectively. β_1 and β_2 : are exponential decay rates. Adam is very tolerant of the initial learning rate (η) and other training parameters. Default values of $\beta_1=0.9$, $\beta_2=0.999$, and $\eta=10^{-8}$ [45].

C. CNN Architecture

We have a dataset of dimensions (32,32,3), and we will perform a conversion from RGB color space to gray level. The input images of our architecture will have dimensions (32,32,1). Table III presented the architecture of CNN in detail, type of layers, output shapes, and activation functions. The layers with their corresponding type are shown, denoting the characteristics used. Then implementation of CNN in CPU takes more time because we have a dataset of images that are more difficult to execute. However, the faster implementation we propose to use GPU.

TABLE III. PROPOSED CNN ARCHITECTURE

Layer	Type	Output shape	param	Activation
Conv2d_1	Conv2D	(None, 28, 28, 60)	1560	ReLU
Conv2d_2	Conv2D	(None, 24, 24, 60)	90060	ReLU
Max_pooling2d_1	Max_pooling2d	(None, 12, 12, 60)	0	N/A
Conv2d_3	Conv2D	(None, 10, 10, 30)	16320	ReLU
Conv2d_4	Conv2D	(None, 8, 8, 30)	8130	ReLU
Max_pooling2d_2	Max_pooling2d	(None, 4, 4, 30)	0	N/A
Dropout_1	Dropout	(None, 4, 4, 30)	0	N/A
Flatten_1	Flatten	(None, 480)	240500	N/A
Dense_1	Dense	(None, 500)	0	ReLU
Dropout_2	Dropout	(None, 500)	21543	N/A
Dense_2	Dense	(None, 43)		Softmax

Total params: 378,023
Trainable params: 378,023
NON TRAINABLE PARAMS: 0

V. EXPERIMENTATION AND EVALUATION

The results are implemented in ASUSTek Computer, processor intel® Core™ i7-7500 CPU @2.70GHz 2.90GHz, Memory installed (RAM): 8,00 Go, exploitation System 64 bits, processor 64 bits Systems Model: X541UJ, GPU NVIDIA GeForce 920M using the TensorFlow, Keras, and OpenCV libraries.

A. Simple Fit Function

The training of the proposed CNN model requires two essential elements, the training data, and the training labels. For the training, we will use the fit function of the Keras library.

The number of epochs is the number of times the model will run through the data. The more epochs we run, the more the model will improve, up to a certain point. We started our model for 50 epochs with a batch size set to 32. We will also train the dataset with equalization and normalization of the histogram. Thus, the training without equalization and normalization will be noted as Method 1, and the training with equalization and normalization will be noted as Method 2.

We can visualize in Fig. 3 in the accuracy curve, a drop during the training of the data in 2 steps for 50 and 100 epochs. But for the loss curve, we have a huge increase in the error value. So, method (1) leads us to overfit. We can deduct from Fig. 4 that we don't have any underfitting or overfitting in the accuracy curve, we can easily observe that the increase in the number of epochs did not disturb the learning stability. The same thing for the loss curve, we have a very remarkable degradation of the error values compared to the curve of

method (1). Equalization and normalization can be used almost. However, this method (2) shows negligible effect loss and we have the full precision of our network that shows a significant improvement.

A comparison of the performance in Table IV shows accuracy function and loss function. We can conclude from Table IV which contains tests accuracy and loss for Method (1) and Method (2). It is necessary to equalize and normalize. The equalization is served to adjust the contrast in the image's dataset. For the normalization, it allows making training faster and the loss becomes more circular symmetric. The next step is to change the simple fit function by using a fit generator, we visualize if does good predictions and evolution of accuracy with a loss function. The equalization and normalization algorithms result in improved performance of CNN classification.

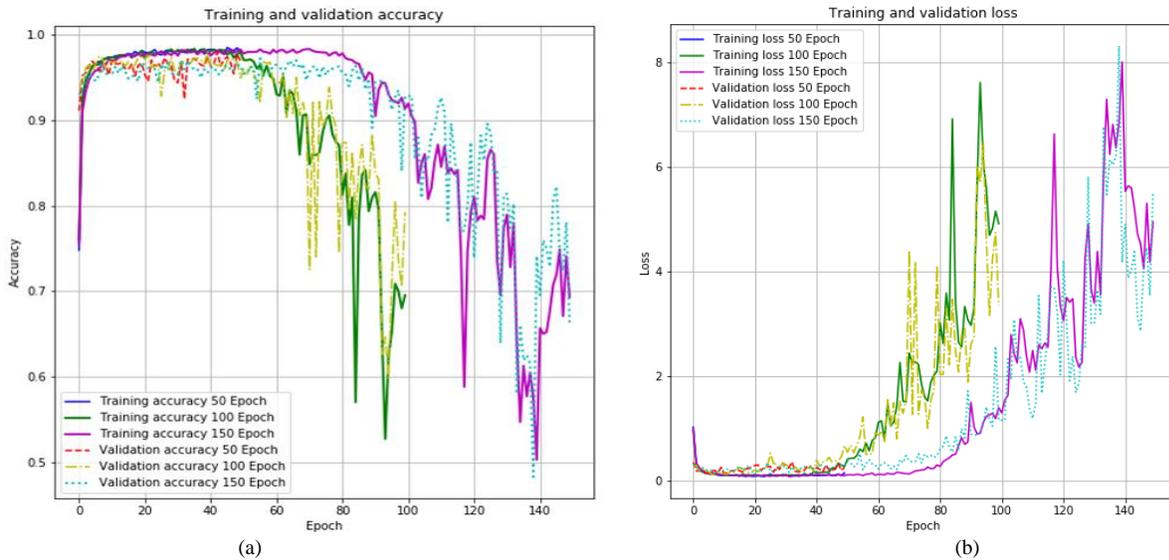


Fig. 3. Method (1): (a) Training and Validation Accuracy, (b) Training and Validation Loss.

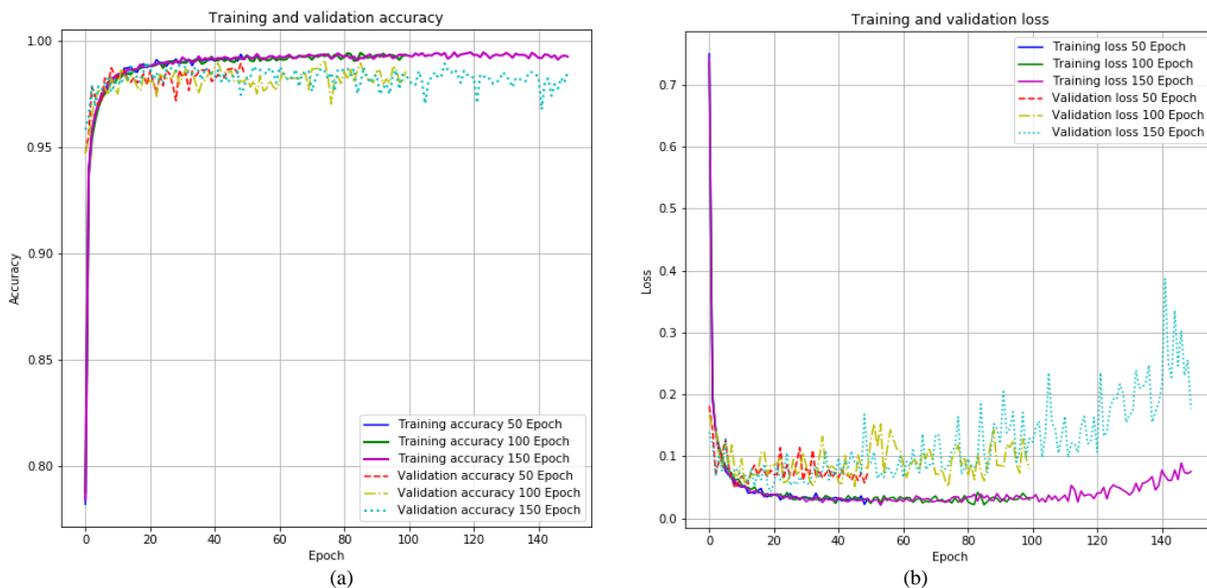


Fig. 4. Method (2): (a) Training and Validation Accuracy, (b) Training and Validation Loss.

TABLE IV. PERFORMANCE COMPARISON OF METHOD 1 AND METHOD 2

Method		Model	Fit Function	Learning Rate	Loss Function	Optimizer	Train Dataset	Epochs	Test Accuracy (%)	Test Loss (%)
Method 1	Without Equalization and Normalization images Datasets	CNN	Simple	10^{-2}	Categorical Cross-Entropy	Adam	34799	50	94.63	44.11
								100	78.21	35.09
								150	64.18	5.76(>100%)
Method 2	With Equalization and Normalization images Datasets							50	96.43	17.65
								100	96.19	25.48
								150	96.52	42.29

B. Fit Generator Function

We propose to use the fit generator function to accept the data sets, perform backpropagation, and update the weights in our model. This function has a hyperparameter, it is the number of steps per epoch, its value as the set of servant landmarks becomes divided by the batch size. It is based on an infinite loop, which must not return empty or exit. However, all researchers calculate the value of steps per epoch as the total number of training data divided by the batch size of training data images.

So, the idea of our experiment is to use method 2 from the previous section. Method 2 will be driven by the generator fitting function with a batch size of 32. We will compare different optimizers (Adam and SGD) and the loss function (Categorical cross-entropy, and Mean squared error). We fixed parameters learning rate in 10^{-2} and epochs in 50.

In Table V, when the loss function uses categorical cross-entropy, we have a high prediction score with a low loss score. Now we improved the model to get the lowest loss score. We got the best scores with the Adam optimizer and the categorical cross-entropy function, for 97.11% accuracy and 11.32% loss. Moreover, the idea is now to improve the accuracy score.

As we can see in Fig. 5, using the fit generator function in the training model the objective is achieved, at 90% we control the situations for not have the overtraining our DL models. The assumptions are therefore correct, we using all of the datasets

at each epoch. We need to choose a batch size and steps per epoch which multiply to give a total number of samples. Usually, it will be a resource. If memory is a problem, we need to reduce the batch size until we can adapt a batch on a GPU. Note that this implementation also allows us to use the multiprocessing argument of a fit generator, where the number of threads specified in workers corresponds to those which generate batches in parallel. A fairly high number of workers ensure that the calculations performed on the GPU are managed efficiently, or in other words, the bottleneck of the whole training process will be due to the propagation operations. In our case, we would probably set batch size the desired amount; we change it only if you want the model to not use all the data for each epoch which deflects the definition of the word epoch.

TABLE V. EXPERIMENT RESULTS OF OPTIMIZER AND LOSS FUNCTION

Loss function	Model	Fit Function	Optimizer	Test Accuracy (%)	Test Loss (%)
Categorical Cross-Entropy	CNN	Generator	SGD	86.74	46.27
			Adam	97.11	11.32
Mean Squared Error			SGD	01.16	02.27
			Adam	97.08	12.12

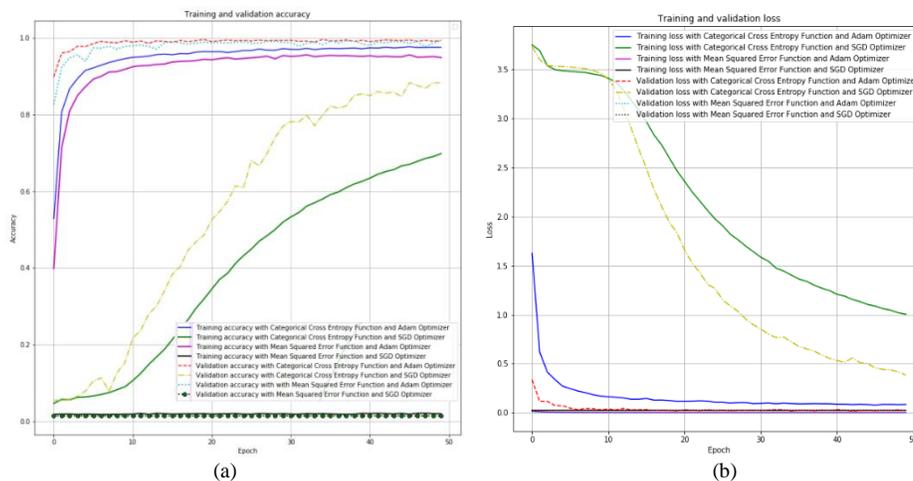


Fig. 5. (a) Training and Validation Accuracy, (b) Training and Validation Loss.

VI. DISCUSSION

We introduce one more technique to improve the model training process data augmentation. This technique creates new data for our CNN model to use during the training process. This is done by taking our existing datasets and transforming or altering the images in useful ways to create new images.

A. Image Data Generator Function

We can have a typical sign image such as this STOP sign image, taking this image and transforming it to create a different image representing the same stop sign. The transformation could be rotation or simply zooming into the image. Also, could even be a combination of both these transformations. These newly created images are referred to as augmented images because they essentially allow us to augment our dataset by adding them. The data augmentation technique is useful because it allows our model to look at each image in our dataset from a variety of different perspectives. This allows it to extract relevant features more accurately and allows it to attain more feature-related data from each training image. This is especially the case for our traffic sign datasets because we have a small dataset (32x32) and a large number of classes. This means that certain classes have very few proximately only 200 training in the Fig. 2. It can benefit our traffic sign recognition model.

We apply the five following transformations with shift range, height shift range, zoom range share, and a rotation range. Five transformations will add sufficient variety to GTSRB datasets and will allow the training process to be much more effective. The first transformation is with shifts, this refers to a horizontal translation in the image which will cause our images to be centered, and this will help our CNN model adapt to test images that aren't necessarily going to be centered. The range can be defined in two ways, if the range value is defined as a number between 0 and 1, then it refers to the fraction of the image that can be shifted. A value of 0.1 would simply imply that the maximum horizontal shift possible is 10 percent of the width of the image. The images with only horizontal translation can be similar. So, to have a difference between the generated, we apply a second technique is a vertical translation. The range value is defined in much the same way and for that reason; the value of vertical translation is 0.1 (10%).

For zoom transformation, can be either zoom out or into the image. The degree of zoom can be defined with a float value between 0 and 1. While the maximum outer zoom is defined by one minus the float value and the maximum zoom is defined by a 1 plus the flow value. We will use a float value of 0.1 which means that we can zoom as far as 0.1 eight's and zoom in as close as 0.2. Next, we have the shear transformation in plane geometry a shear mapping is a linear map that displaces, each point in a fixed direction by an amount proportional to its side and distance. The line that is parallel to that direction, possible in both the x-direction and the y-direction. This transformation is defined using shear intensity which simply refers to the magnitude of the shear, angle in degrees as seen in the image above. We apply a small magnitude of shear to be effective, using a value of 0.1. The last transformation is the rotation; this transformation is a bit more intuitive it simply rotates an image

by a certain value of degrees. This value can be defined using an integer value, in our case, we will use 10. These transformations are simply applied to stop signs as shown in Fig. 6, which will then be applied to the GTSRB dataset.

B. Batch Size Function

First, we declare a batch size is equal to 32 which mean that our image generator will create a batch of 32 images at a time for our CNN model to use our next argument as illustrated in Fig. 7. Also, the steps per epoch this parameter essentially refers to the number of batches. The steps per-epoch argument must specify the number of batches of samples comprising one epoch. In our case, the original dataset has 34799 images and the batch size is 32. Then a reasonable value for steps per epoch when fitting a model on the augmented data might be $\text{ceil}(34799/32)$, or 1087 batches. So, we fix the value of the steps per epoch in 1000.

C. Experimental Results

We are fixed step pre-epochs to 1000, we switch the value of epochs between 50 and 150, we behold augmentation accuracy the same time value loss has diminution. We fit and evaluate all these models in different batch sizes (32, 64, 128, and 256) using the same procedure above of optimizer Adam and the same value of steps pre-epochs with different epochs, found through some minor experimentation. The model is evaluated, reporting the classification accuracy on the test sets between 96.86% and 98.01%. We can specify the results may vary given the stochastic nature of the training algorithm. Table VI demonstrates the effect of batch size, after testing very hard which took an enormous time to find it up to incredible values. When we have for batch size is 256, we have a precision in 50 epochs of 98,01% which is very interesting, and also a remarkable reduction in the function error of 09.15%. The same thing for size 100 epochs has values for the two 97.99% and 09.11%.

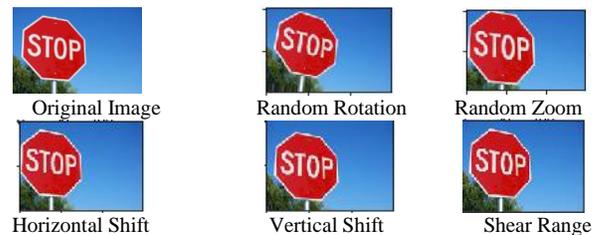


Fig. 6. Different Transformation for Dataset Augmentation.

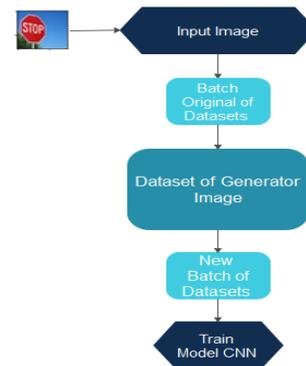


Fig. 7. The Batch of the Training Dataset GTSRB.

TABLE VI. FLOW DATA FOR BATCH SIZE FUNCTION

Batch Size	Optimizer	Step pre-epochs	Epoch	Test Accuracy (%)	Test Loss (%)
32	Adam	1000	50	97.15	10.94
			100	97.11	11.32
			150	96.86	14.16
64	Adam	1000	50	97.18	11.90
			100	97.04	12.28
			150	97.69	09.25
128	Adam	1000	50	97.38	11.02
			100	97.85	10.74
			150	97.94	09.68
256	Adam	1000	50	98.01	09.15
			100	97.99	09.11
			150	97.83	10.84

In Fig. 8, we can see the validation converges to above 99%. A significant improvement is shown over our previous CNN model. This might be our modification that was pretty effective. We have a much smaller gap and training accuracy as well as our validation loss and accuracy, respectively. This demonstrates consistency in our training and a better-trained model and we now finish our model training with a validation accuracy of over 98 % and training accuracy. This is all very good to see and shows our augmentation technique was effective. The model will not learn complex patterns and we can avoid overfitting, we use more dropout layers in our architecture and check its performance. So, the augmentation dataset after performing histogram normalization and equalization, the model learned the data better, and the accuracy of the set improved. Now there is just one more test that our model needs to pass and that is classifying images from the test dataset to predict a couple of them correctly. So, we'll start by testing out the image not seen before for our CNN model.

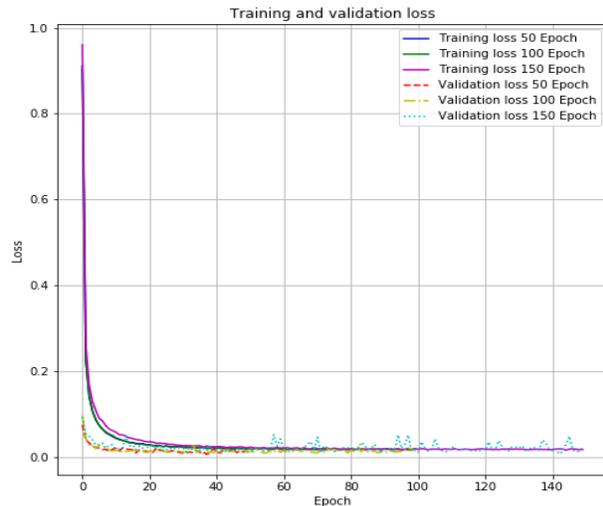
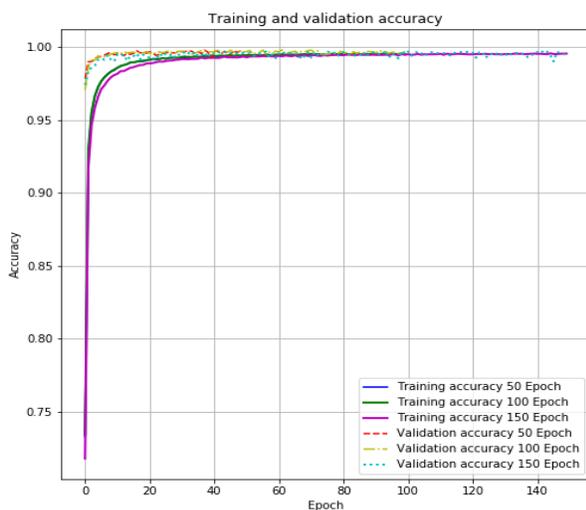


Fig. 8. Accuracy and Loss Curves of Epochs with a Batch Size of 256.

D. Analyses Performance of a Model Trained

We define several measures based on the confusion matrix, to quantify the performance of a classifier from different points of view: Precision by class, average precision, Recall by class, average recall, F-score by class, and f-score average.

a) *Precision of classification*: The accuracy of a classifier concerning a certain class in other words, about a certain modality of the variable to be predicted, is measured as the proportion of individuals, among all those for whom the classifier predicted this class, who belong to it, exposed in this equation:

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

TP : (True Positive) Element of the class correctly predicted.

FP: (False Positive) Element of the class badly predicted.

The overall means of the precision over all the classes i can be evaluated by the macro-average which first calculates the precision on each class i followed by a calculation of the average of the details on the n classes based on this equation:

$$Precision = \sum_i^n \frac{p_i}{n} \quad \# \quad (9)$$

p_i : precision each class i. n : number of classes.

b) *Recall of classification*: The recall of a classifier with a certain class is measured, like the proportion of individuals, among all those who belong to this class, for which the classifier predicted this class.

$$r_i = \frac{TP}{TP+FN} \quad \# \quad (10)$$

The global averages of the recall over all of the classes i can be evaluated by the macro-average which first calculates the recall over each class i followed by a calculation of the average of the reminders over the n classes:

$$Recall = \sum_i^n \frac{r_i}{n} \quad \# \quad (11)$$

r_i : recall each class i. n : number of classes.

c) *F-score of classification*: We can summarize the recall precision measurements to a class in a single indicator, by calculating the harmonic mean:

$$RF - score_i = \frac{2*(p_i*r_i)}{p_i+r_i} \quad \# \quad (12)$$

p_i : precision each class i. r_i : recall each class i.

The average over each class of these indicators gives global indicators on the quality of the classifier.

$$F_{-Score} = \frac{2*(Precision*Recall)}{Precision+Recall} \quad \# \quad (13)$$

Precision: The average precision of all classes.

Recall: The average recall of all classes.

E. *Confusion Matrix*

The Confusion Matrix identifies the classes of signs and also gives the number of times it gets the confused class to identify the class from another in Fig. 9. Most of the color is diagonal, but there are still some annoying spots somewhere. When we narrowly look at the confusion matrix, we see that the classes [0] have very less respectively all classes, but it's minimized for other classes. The diagonal observations are the true positives of each class and other non-diagonal observations are incorrect classifications of the model.

F. *Classifier Metrics*

A Classification report is used to measure the quality of predictions from a classification algorithm. We can see in the Table VII, the model has the as recall and precision are calculated for individual classes, have a good score of all the class of traffic road signs. We use macro or micro or weighted scores of recalls, precision, and F1 score of a model for multiclass classification problems have a higher score is 98% this very satisfied.

TABLE VII. CLASSIFIER REPORT FOR THE CNN MODEL

Class names	precision	recall	f1-score	support
Speed limit (20km/h)	0.98	1.00	0.99	60
Speed limit (30km/h)	0.99	1.00	0.99	720
Speed limit (50km/h)	0.99	0.99	0.99	750
Speed limit (60km/h)	0.99	0.94	0.96	450
Speed limit (70km/h)	1.00	0.98	0.99	660
Speed limit (80km/h)	0.95	0.99	0.97	630
End of speed limit (80km/h)	0.99	0.90	0.94	150
Speed limit (100km/h)	1.00	1.00	1.00	450
Speed limit (120km/h)	1.00	1.00	1.00	450
No passing	1.00	1.00	1.00	480
No passing for vechiles over 3.5 metric tons	1.00	1.00	1.00	660
Right-of-way at the next intersection	0.98	0.96	0.97	420
Priority road	1.00	0.99	1.00	690
Yield	1.00	0.99	1.00	720
Stop	1.00	0.99	0.99	270
No vechiles	1.00	0.97	0.98	210
Vechiles over 3.5 metric tons prohibited	0.99	1.00	1.00	150
No entry	1.00	0.97	0.99	360
General caution	0.99	0.89	0.94	390
Dangerous curve to the left	0.98	1.00	0.99	60
Dangerous curve to the right	0.98	1.00	0.99	90
Double curve	0.85	0.78	0.81	90
Bumpy road	1.00	0.93	0.96	120
Slippery road	0.98	1.00	0.99	150
Road narrows on the right	0.99	0.98	0.98	90
Road work	0.98	0.97	0.97	480
Traffic signals	0.91	0.98	0.95	180
Pedestrians	0.89	0.95	0.92	60
Children crossing	0.99	1.00	1.00	150
Bicycles crossing	1.00	0.99	0.99	90
Beware of ice/snow	0.89	0.95	0.92	150
Wild animals crossing	1.00	1.00	1.00	270
End of all speed and passing limits	1.00	0.98	0.99	60
Turn right ahead	0.96	1.00	0.98	210
Turn left ahead	0.99	1.00	1.00	120
Ahead only	1.00	1.00	1.00	390
Go straight or right	0.99	1.00	1.00	120
Go straight or left	0.97	0.98	0.98	60
Keep right	1.00	0.97	0.98	690
Keep left	1.00	0.98	0.99	90
Roundabout mandatory	0.85	0.93	0.89	90
End of no passing	0.98	1.00	0.99	60
End of no passing by vechiles over 3.5 metric tons	0.95	0.87	0.91	90
micro avg	0.98	0.98	0.98	12630
macro avg	0.98	0.97	0.97	12630
weighted avg	0.99	0.98	0.98	12630
samples avg	0.98	0.98	0.98	12630

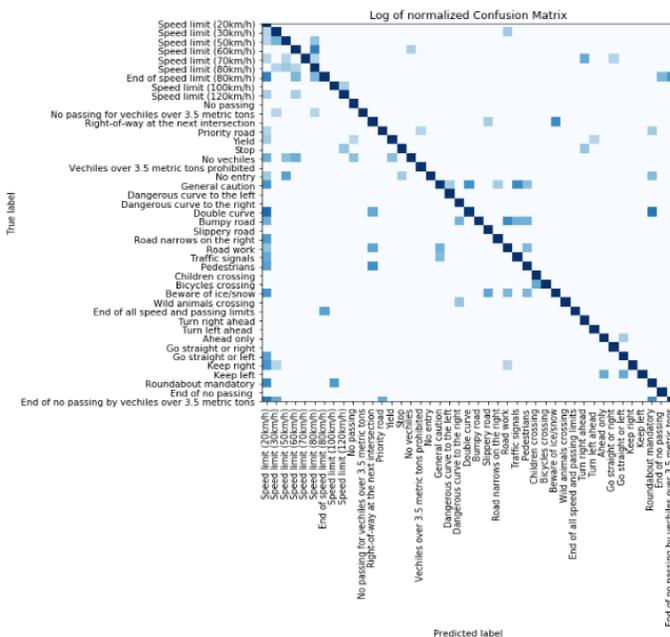


Fig. 9. Confusion Matrix Epochs 100 and Batch Size 256.

VII. TESTING THE MODEL

A. Test with the Test Dataset

A remarkable performance is illustrated in Fig. 10. Now we'll test for test datasets, we look reaction to our model, to see where the fails. We tried to visualize the class predictions of the test images, it is relevant to have good results, all the images were well classified, and the curve shown next to each image represents the class of the images among the 42 classes, when we have the color blue and a single peak in the curve means the image has been put in the right place without any errors.

B. Testing the Proposed CNN Model in Real-Time

In this section, we will present and evaluate the results of our approach. Traffic road signs that appear in video sequences are often detected. More details on the video sequences are given in Table VIII. In general, for all performance indicators, our proposed approach outperforms other object detection algorithms by achieving up to 100% accuracy. Our CNN model metric value is often higher than in the results of previous work. For the video sequences, our algorithm surpasses the good probability of prediction and classes of Traffic Road signs by the method. This shows that using a robust appearance CNN model achieves better results. It can also be observed that the CNN precision value obtained for the video sequence is higher than that obtained by the approach with a difference between 97.56% and 100%.

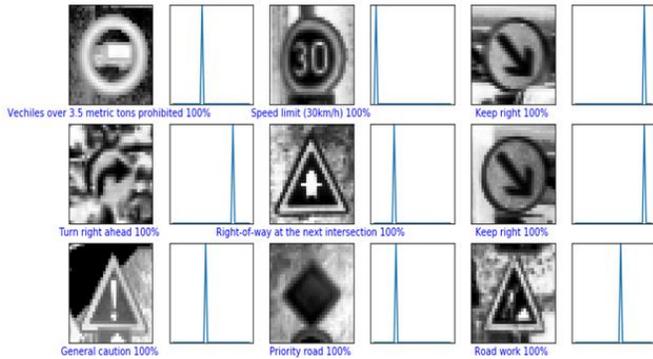


Fig. 10. GRTSB Test Datasets.

TABLE VIII. TEST OF NEW IMAGES IN REAL-TIME

Traffic Road Signs Recognition	Classification	Prediction (%)
	Class Number: [17] No Entry	97.56%
	Class Number: [12] Priority Road	99.15%
	Class Number: [5] Speed limit (80km/h)	98.4%
	Class Number: [14] STOP	100%
	Class Number: [34] Turn let ahead	100%
	Class Number: [33] Turn right ahead	

VIII. CONCLUSION

In this paper, we proposed a methodology for the construction robust CNNs model. We talked about the problems associated with the detection and recognition of traffic road signs in real-time. We also demonstrated how using the right tools and techniques helps us in developing robust CNN models. These CNNs can guarantee road safety in real-time. We also try other pre-processing techniques to further improve the model's accuracy (equalization and normalization histogram). The step of adding augmentation data improved the performance of our deep learning CNN model. We are curious about how much the accuracy can be improved based on adding such simple transformations. We think these results could further be used in the development of automotive systems, such as intelligent transportation systems (ITS). All this is for the safest roads; we try in the future to get better performance and optimistic. It is also very interesting to note that the proposed CNN model reaches 98% accuracy using NVIDIA's GPU processor, which makes them feasible for real-time traffic sign recognition.

In future work, we plan to study other neural network architectures that have been shown to be optimistic for traffic sign detection or classification. In addition, we will attempt to employ these networks in advanced in-vehicle platforms applicable to intelligent transportation systems to reveal valuable information that will help drivers make the right decisions in the real world.

REFERENCES

- [1] X. Xu, J. Jin, S. Zhang, L. Zhang, S. Pu, and Z. Chen, "Smart data driven traffic sign detection method based on adaptive color threshold and shape symmetry," *Futur. Gener. Comput. Syst.*, vol. 94, pp. 381–391, 2019, doi: 10.1016/j.future.2018.11.027.
- [2] S. Kumar, S. D. K. P. Maddula, and N. V. V. Ravipati, "Unified approach for detecting traffic signs and potholes on Indian roads," *J. King Saud Univ. - Comput. Inf. Sci.*, no. xxxx, 2021, doi: 10.1016/j.jksuci.2021.12.006.
- [3] D. Sirohi, N. Kumar, and P. Singh, "Convolutional neural networks for 5G-enabled Intelligent Transportation System: A systematic review," *Comput. Commun.*, vol. 153, no. January, pp. 459–498, 2020, doi: 10.1016/j.comcom.2020.01.058.
- [4] S. L. G. Eliza, "Embedded real-time speed limit sign recognition using image processing and machine learning techniques," vol. 28, pp. 573–584, 2017, doi: 10.1007/s00521-016-2388-3.
- [5] A. Zemmouri, M. Alareqi, R. Elgouri, M. Benbrahim, and L. Hlou, "Integration and implementation system-on-a-programmable-chip (SOPC) in FPGA," *J. Theor. Appl. Inf. Technol.*, vol. 76, no. 1, pp. 127–133, 2015.
- [6] S. Jagannathan, M. Mody, and M. Mathew, "Optimizing Convolutional Neural Network on DSP," pp. 371–372, 2016.
- [7] A. Zemmouri, R. Elgouri, M. Alareqi, M. Benbrahim, and L. Hlou, "Design and implementation of pulse width modulation using hardware/software microblaze soft-core," *Int. J. Power Electron. Drive Syst.*, vol. 8, no. 1, pp. 167–175, 2017, doi: 10.11591/ijpeds.v8i1.pp167-175.
- [8] W. Arman, S. Arefin, A. S. M. Shihavuddin, and M. Abul, "DeepThin : A novel lightweight CNN architecture for traffic sign recognition without GPU requirements," *Expert Syst. Appl.*, vol. 168, no. August 2020, p. 114481, 2021, doi: 10.1016/j.eswa.2020.114481.
- [9] A. Zemmouri, R. Elgouri, M. Alareqi, H. Dahou, M. Benbrahim, and L. Hlou, "A comparison analysis of PWM circuit with arduino and FPGA," *ARPN J. Eng. Appl. Sci.*, vol. 12, no. 16, pp. 4679–4683, 2017.
- [10] Y. Saadna, "Speed limit sign detection and recognition system using SVM and MNIST datasets," *Neural Comput. Appl.*, vol. 0123456789, 2019, doi: 10.1007/s00521-018-03994-w.
- [11] A. R. Dubey, N. Shukla, and D. Kumar, "Detection and Classification of Road Signs Using HOG-SVM Method," pp. 49–56.
- [12] Z. Liu, M. Qi, C. Shen, Y. Fang, and X. Zhao, "Cascade saccade machine learning network with hierarchical classes for traffic sign detection," *Sustain. Cities Soc.*, vol. 67, no. January, p. 102700, 2021, doi: 10.1016/j.scs.2020.102700.
- [13] C. Li, Z. Qu, S. Wang, and L. Liu, "A method of cross-layer fusion multi-object detection and recognition based on improved faster R-CNN model in complex traffic," *Pattern Recognit. Lett.*, vol. 145, pp. 127–134, 2021, doi: 10.1016/j.patrec.2021.02.003.
- [14] H. Li et al., "A defense method based on attention mechanism against traffic sign adversarial samples," *Inf. Fusion*, vol. 76, no. March 2020, pp. 55–65, 2021, doi: 10.1016/j.inffus.2021.05.005.
- [15] H. Kwon et al., "NeuroImage Early cortical signals in visual stimulus detection," *Neuroimage*, vol. 244, no. September, p. 118608, 2021, doi: 10.1016/j.neuroimage.2021.118608.
- [16] S. Messaoud, S. Bouaafia, A. Maraoui, and A. Chiheb, "Deep convolutional neural networks-based Hardware – Software on-chip system for computer vision application ☆," *Comput. Electr. Eng.*, vol. 98, no. June 2021, p. 107671, 2022, doi: 10.1016/j.compeleceng.2021.107671.
- [17] E. Karaaslan, U. Bagci, and F. N. Catbas, "Attention-guided analysis of infrastructure damage with semi-supervised deep learning," *Autom. Constr.*, vol. 125, no. April 2019, p. 103634, 2021, doi: 10.1016/j.autcon.2021.103634.
- [18] K. Nogueira, O. A. B. Penatti, and J. A. dos Santos, "Towards better exploiting convolutional neural networks for remote sensing scene classification," *Pattern Recognit.*, vol. 61, pp. 539–556, 2017, doi: 10.1016/j.patcog.2016.07.001.
- [19] Y. Anagun, S. Isik, and E. Seke, "SRLibrary: Comparing different loss functions for super-resolution over various convolutional architectures," *J. Vis. Commun. Image Represent.*, vol. 61, pp. 178–187, 2019, doi: 10.1016/j.jvcir.2019.03.027.
- [20] K. Fu, Q. Zhao, I. Yu-Hua Gu, and J. Yang, "Deepside: A general deep framework for salient object detection," *Neurocomputing*, vol. 356, pp. 69–82, 2019, doi: 10.1016/j.neucom.2019.04.062.
- [21] Y. Wang, Z. Fang, and H. Hong, "Comparison of convolutional neural networks for landslide susceptibility mapping in Yanshan County, China," *Sci. Total Environ.*, vol. 666, pp. 975–993, 2019, doi: 10.1016/j.scitotenv.2019.02.263.
- [22] B. Zhao, X. Li, X. Lu, and Z. Wang, "A CNN-RNN architecture for multi-label weather recognition," *Neurocomputing*, vol. 322, pp. 47–57, 2018, doi: 10.1016/j.neucom.2018.09.048.
- [23] J. Du, C. Vong, S. Member, and C. L. P. Chen, "Novel Efficient RNN and LSTM-Like Architectures: Recurrent and Gated Broad Learning Systems and Their Applications for Text Classification," pp. 1–12, 2020.
- [24] H. Chen, L. Wu, J. Chen, W. Lu, and J. Ding, "A comparative study of automated legal text classification using random forests and deep learning," *Inf. Process. Manag.*, vol. 59, no. 2, p. 102798, 2022, doi: 10.1016/j.ipm.2021.102798.
- [25] Z. Parcheta, G. Sanchis-Trilles, F. Casacuberta, and R. Rendahl, "Combining Embeddings of Input Data for Text Classification," *Neural Process. Lett.*, vol. 53, no. 5, pp. 3123–3151, 2021, doi: 10.1007/s11063-020-10312-w.
- [26] X. Li, M. Cui, J. Li, R. Bai, Z. Lu, and U. Aickelin, "A hybrid medical text classification framework: Integrating attentive rule construction and neural network," *Neurocomputing*, vol. 443, pp. 345–355, 2021, doi: 10.1016/j.neucom.2021.02.069.
- [27] M. A. Ibrahim, M. U. Ghani Khan, F. Mehmood, M. N. Asim, and W. Mahmood, "GHS-NET a generic hybridized shallow neural network for multi-label biomedical text classification," *J. Biomed. Inform.*, vol. 116, no. April 2020, p. 103699, 2021, doi: 10.1016/j.jbi.2021.103699.
- [28] F. Nielsen, "Anticipation-RNN: enforcing unary constraints in sequence generation, with application to interactive music generation," vol. 4, 2018, doi: 10.1007/s00521-018-3868-4.

- [29] D. Bisharad, "Music genre recognition using convolutional recurrent neural network architecture," no. April, pp. 1–13, 2019, doi: 10.1111/exsy.12429.
- [30] A. Onan, "Sentiment analysis on massive open online course evaluations: A text mining and deep learning approach," *Comput. Appl. Eng. Educ.*, vol. 29, no. 3, pp. 572–589, 2021, doi: 10.1002/cae.22253.
- [31] J. L. Huan, A. A. Sekh, C. Quek, and D. K. Prasad, "Emotionally charged text classification with deep learning and sentiment semantic," *Neural Comput. Appl.*, vol. 34, no. 3, pp. 2341–2351, 2022, doi: 10.1007/s00521-021-06542-1.
- [32] Á. Arcos-García, J. A. Álvarez-García, and L. M. Soria-Morillo, "Deep neural network for traffic sign recognition systems: An analysis of spatial transformers and stochastic optimisation methods," *Neural Networks*, vol. 99, pp. 158–165, 2018, doi: 10.1016/j.neunet.2018.01.005.
- [33] Á. Arcos-García, M. Soilán, J. A. Álvarez-García, and B. Riveiro, "Exploiting synergies of mobile mapping sensors and deep learning for traffic sign recognition systems," *Expert Syst. Appl.*, vol. 89, pp. 286–295, 2017, doi: 10.1016/j.eswa.2017.07.042.
- [34] J. Li, X. Mei, S. Member, D. Prokhorov, and S. Member, "Deep Neural Network for Structural Prediction and Lane Detection in Traffic Scene," pp. 1–14, 2016.
- [35] J. Kim, J. Kim, G. Jang, and M. Lee, "Fast learning method for convolutional neural networks using extreme learning machine and its application to lane detection," *Neural Networks*, vol. 87, pp. 109–121, 2017, doi: 10.1016/j.neunet.2016.12.002.
- [36] J. Tang, S. Li, and P. Liu, "A review of lane detection methods based on deep learning," *Pattern Recognit.*, vol. 111, p. 107623, 2021, doi: 10.1016/j.patcog.2020.107623.
- [37] A. Shustanov and P. Yakimov, "CNN Design for Real-Time Traffic Sign Recognition," *Procedia Eng.*, vol. 201, pp. 718–725, 2017, doi: 10.1016/j.proeng.2017.09.594.
- [38] F. Shao, X. Wang, F. Meng, T. Rui, D. Wang, and J. Tang, "Real-Time Traffic Sign Detection and Recognition Method Based on Simplified Gabor Wavelets," 2018, doi: 10.3390/s18103192.
- [39] T. Yang, X. Long, A. Kumar, Z. Zheng, and C. Tong, "Deep detection network for real-life traffic sign in vehicular networks," vol. 136, pp. 95–104, 2018, doi: 10.1016/j.comnet.2018.02.026.
- [40] Á. Arcos-García, J. A. Álvarez-García, and L. M. Soria-Morillo, "Evaluation of deep neural networks for traffic sign detection systems," *Neurocomputing*, vol. 316, pp. 332–344, 2018, doi: 10.1016/j.neucom.2018.08.009.
- [41] A. Barodi, A. Bajit, M. Benbrahim, and A. Tamtaoui, "An Enhanced Approach in Detecting Object Applied to Automotive Traffic Roads Signs," in *6th International Conference on Optimization and Applications, ICOA 2020 - Proceedings*, Apr. 2020, pp. 1–6, doi: 10.1109/ICOA49421.2020.9094457.
- [42] U. A. Nnolim, "An adaptive RGB colour enhancement formulation for logarithmic image processing-based algorithms," *Optik (Stuttg.)*, vol. 154, pp. 192–215, 2018, doi: 10.1016/j.ijleo.2017.09.102.
- [43] A. Barodi, A. Bajit, M. Benbrahim, and A. Tamtaoui, "Applying Real-Time Object Shapes Detection to Automotive Traffic Roads Signs," 2020, doi: 10.1109/ISAECT50560.2020.9523673.
- [44] N. Ben Romdhane, H. Mliki, R. El Beji, and M. Hammami, "Combined 2d/3d traffic signs recognition and distance estimation," *IEEE Intell. Veh. Symp. Proc.*, vol. 2016-Augus, no. Iv, pp. 355–360, 2016, doi: 10.1109/IVS.2016.7535410.
- [45] A. Barodi, A. Bajit, A. Tamtaoui, and M. Benbrahim, "An Enhanced Artificial Intelligence-Based Approach Applied to Vehicular Traffic Signs Detection and Road Safety Enhancement," *Adv. Sci. Technol. Eng. Syst. J.*, vol. 6, no. 1, pp. 672–683, 2021, doi: 10.25046/aj060173.
- [46] A. Barodi, A. Bajit, M. Benbrahim, and A. Tamtaoui, "Improving the transfer learning performances in the classification of the automotive traffic roads signs," *E3S Web Conf.*, vol. 234, no. February, 2021, doi: 10.1051/e3sconf/202123400064.
- [47] J. Stallkamp, M. Schlipsing, J. Salmen, and C. Igel, "The German Traffic Sign Recognition Benchmark for the IJCNN'11 Competition," *Proc. Int. Jt. Conf. Neural Networks*, pp. 1453–1460, 2011, [Online]. Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6033395.
- [48] R. Udendhran, M. Balamurugan, A. Suresh, and R. Varatharajan, "Enhancing image processing architecture using deep learning for embedded vision systems," *Microprocess. Microsyst.*, vol. 76, p. 103094, 2020, doi: 10.1016/j.micpro.2020.103094.
- [49] B. S. Rao, "Dynamic Histogram Equalization for contrast enhancement for digital images," *Appl. Soft Comput. J.*, vol. 89, p. 106114, 2020, doi: 10.1016/j.asoc.2020.106114.
- [50] A. Bouti, M. A. Mahraz, J. Riffi, and H. Tairi, "A robust system for road sign detection and classification using LeNet architecture based on convolutional neural network," *Soft Comput.*, vol. 24, no. 9, pp. 6721–6733, 2020, doi: 10.1007/s00500-019-04307-6.
- [51] T. Sercu and V. Goel, "Dense Prediction on Sequences with Time-Dilated Convolutions for Speech Recognition," no. Nips, 2016, [Online]. Available: <http://arxiv.org/abs/1611.09288>.
- [52] H. C. Shin et al., "Deep Convolutional Neural Networks for Computer-Aided Detection: CNN Architectures, Dataset Characteristics and Transfer Learning," *IEEE Trans. Med. Imaging*, vol. 35, no. 5, pp. 1285–1298, 2016, doi: 10.1109/TMI.2016.2528162.
- [53] M. Gao, Q. Zhang, J. Dong, D. Yang, and D. Zhou, "End-to-end speech emotion recognition based on one-dimensional convolutional neural network," *ACM Int. Conf. Proceeding Ser.*, vol. Part F1481, pp. 78–82, 2019, doi: 10.1145/3319921.3319963.
- [54] B. Zoph and Q. V Le, "Searching for activation functions," *6th Int. Conf. Learn. Represent. ICLR 2018 - Work. Track Proc.*, pp. 1–13, 2018.
- [55] 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.

Research on Students' Course Selection Preference based on Collaborative Filtering Algorithm

Mustafa Man¹, Jianhui Xu², Ily Amalina Ahmad Sabri³, Jiaxin Li⁴

Faculty of Ocean Engineering Technology and Informatics, Universiti Malaysia Terengganu, Terengganu, Malaysia^{1,2,3}
Software College, Liaoning Technical University, Anshan, China⁴

Abstract—Due to the events caused by the COVID-19 pandemic, the education industry is no longer limited to offline, and online classroom education is widely used. The rapid development of online education provides users with more abundant educational course resources and flexible learning methods. Various online education platforms are also constantly improving their service models to give users a better learning experience. However, at present, there are few personalized information recommendation services in student course selection. Students receive the same course selection information and cannot be "tailored" according to their specific preferences. This paper focuses on the integration of collaborative filtering technology into a college course selection system to construct a rating matrix based on students' ratings of the courses they take through correlation between courses and correlation between students. Based on the collaborative filtering algorithm, a predictive rating matrix is generated to produce a recommendation list to achieve intelligent recommendation of suitable courses for students. The experimental results show that, based on the traditional collaborative filtering recommendation technique, the improved collaborative filtering algorithm based on both item and user weighting is used to achieve course recommendation with higher recommendation accuracy. The application of the improved collaborative filtering technique in the course selection recommendation system of colleges and universities is very good at recommending courses for students intelligently, and the recommended courses for students have good rationality and accuracy, and achieve more intelligent course selection for students, which has great practicality and practical significance.

Keywords—Collaborative filtering; course selection system; recommendation; scoring matrix; weighting

I. INTRODUCTION

With the rapid development of the Internet industry, the application of information technology is becoming more and more widespread in the management of academic affairs information, and the online course selection system has become a significance part of the management of academic affairs information. How to quickly find out the course you are interested in among the large amount of optional course information has become one of the research hot-spots of the course selection system. In the course selection system, most of them are based on the system search engine to query course information and take courses [1]. When facing a large number of available courses, students do not have the relevant knowledge as a basis to select courses because they do not know enough about the course information, which will lead to a waste of course resources. Currently, most of the course

selection systems have no recommendation function or low quality of personalized recommendation, so it does not recommend the courses that students may be interested in. However, building a recommendation system based on feedback information such as students' major information and students' ratings of course information can solve these problems [2].

A. The Statement of the Research Problem

Elective courses for college students are courses that students can choose to study independently according to their preferences. Through elective courses, students can expand their knowledge. At present, the course selection system of most colleges and universities lists all the elective courses offered this semester in the system like commodities for students to choose [3]. The relevance and orientation between courses are poor. When college students choose courses, there are the following three problems:

1) *Blindness*: Students do not understand the relevance of courses and the direction of majors, Course selection is arbitrary.

2) *Poor purpose, dealing with errands*: Students only choose some courses that can easily pass in order to complete their credits, regardless of whether the courses they learn are helpful to their curriculum system. As a result, after the course is selected, the learning enthusiasm is poor and the learning effect is not good, which does not achieve the expected effect of elective courses (R. N. Behera and S. Dash, 2016).

3) *Instability and potential risks of course selection system*: The traditional course selection mode has strong time constraint and does not take into account algorithmic fairness, which may often cause peak access and easily cause hidden danger to the security operation of the back-end system.

Therefore, personalized recommendation technology is applied to the course selection system to provide students with personalized elective course recommendations according to their needs and interest preferences, prevent students from choosing courses blindly, and greatly improve the utilization rate of university elective course resources and the operation efficiency of the course selection system.

B. Research Objectives

1) To analyzes and compares several recommendation algorithms, finds out their shortcomings and advantages, and determines the research idea of applying collaborative recommendation technology to course selection system.

2) By understanding the courses that students have taken and their evaluations, analyzing students' preferences, and pushing courses that target students may be willing to take.

3) To apply collaborative filtering technology to the field of students' course selection, and combine it with the basic course selection system to realize an efficient and convenient Personalized Course Selection recommendation system.

C. Research Question

The recommendation algorithm based on collaborative filtering is based on the similarity measure of individual behavioral characteristics. By calculating the similarity between the specified new sample (students) and the original sample in the database, the new sample is clustered, and the individuals with similar behavioral characteristics to the new sample are identified as the nearest neighbor samples (nearest students). After that, the selection set of the nearest neighbor sample is generated, and the selection set is sorted according to preference scores.

Ultimately, course selection recommendations are made to students based on the similarity of the new sample to the nearest neighbor students and the course selection preferences of the nearest neighbor students. There are two important issues that need to be addressed.

1) First, how to calculate the similarity between students through behavioral data, and require that the similarity reflect the interests and learning characteristics of students strengths.

2) Second, after identifying a new sample of near-neighbor students, how to determine the set of recommended courses to be selected based on the near-neighbor students' course selection records.

D. Rationale of the Study

In the field of teaching management practice of colleges and universities, in order to follow up the reform of higher education teaching and meet the demand of society for comprehensive and practical talents, the course selection system of China's colleges and universities has completed the conversion from academic year system to credit system, and schools provide students with a large number of elective courses, and students in colleges and universities are given more options to choose their favorite courses according to their interests. With the diversified development of the society, people are more and more concerned with a wide range of fields, and students' interests show a trend of divergence, so colleges and universities have also opened corresponding courses for students to choose in response to this phenomenon. However, in recent years, there are too many elective courses in colleges and universities, which lead to information overload when students choose courses, and it is difficult for students to choose courses that suit their personal development due to the structural deficiencies of course classification and specialization. This shows that the course selection process of students is not without research value, and there is a pattern of course selection behavior. Moreover, the traditional course selection mode has strong time constraint and does not take into account algorithmic fairness, which may often cause peak access and easily pose hidden risks to the security operation of

the back-end system. Therefore, personalized recommendation technology is applied to the course selection system to provide students with personalized elective course recommendations according to their needs and interest preferences, prevent students from choosing courses blindly, and greatly improve the utilization rate of university elective course resources and the operation efficiency of the course selection system [4].

The "department store" approach of simply improving the quality of teaching resources by simply listing them and letting students "pick and choose" is obviously no longer in line with the current requirements for "personalized learning".

Personalized learning requires adding a number of "shoppers" to the "department store" with a wide range of elective courses to help learners get the right course for them in a timely and accurate manner that is recognized by the learners. This "shopper" is the role that learning paths play in the learning process of learners, aiming to improve the precise guidance of learners, reduce the blindness of learners, and improve the efficiency of course selection [5].

E. Research Gap

At the same time, with the digital reform of higher education, some scholars have started to study the mining of student's one-card data to analyze student behavior. The students' one-card accumulates a large amount of student spending data and daily behavior data. However, there are still many shortcomings in the above-mentioned research on course selection recommendation systems. The recommendation systems based on students' course selection data use very limited course selection data of varying quality, which makes it difficult to accurately mine students' preferences, and they focus too much on the algorithm level, trying to copy the success of recommendation systems in e-commerce and entertainment fields to the education field, using various methods to improve the accuracy of the algorithm, while ignoring the characteristics of the education field itself and the limitations of the scoring matrix itself, resulting in no qualitative improvement in accuracy and hardly satisfactory recommendation results.

Web mining and bibliography mining are the theoretical basis of data mining technology in library user behavior analysis. Based on the research and analysis of library patron behavior composition and acquisition, library user behavior models are constructed by using machine learning and other algorithms. Through these models, we can understand the interest preferences of reader groups. However, there are still many shortcomings in the above-mentioned research on course selection recommendation systems. The recommendation systems based on students' course selection data use very limited course selection data of varying quality, which makes it difficult to accurately mine students' preferences, and they focus too much on the algorithm level, trying to copy the success of recommendation systems in e-commerce and entertainment fields to the education field, using various methods to improve the accuracy of the algorithm, while ignoring the characteristics of the education field itself and the limitations of the scoring matrix itself, resulting in no qualitative improvement in accuracy and hardly satisfactory recommendation results.

This paper proposes a collaborative filtering algorithm-based course selection recommendation system, which no longer pursues excessive algorithmic complexity, avoids the limitations of the scoring matrix itself, and realizes personalized course recommendation.

II. LITERATURE REVIEW

Collaborative filtering is a push technology that is often used to achieve the basic recommendation push function for the system. It is mainly to divide users into different sets by different tendencies, and to push items to target users according to the items that are closer to users' preferences, item's comment information, etc. as the basis for judging.

A. Theoretical Background

Collaborative filtering is to mine a small number of students with similar course preferences to the specified students in a large amount of data, and then designate these similar students as Then, we organize the course preferences of the near-neighbor students into a catalog sorted by preference, and finally recommend courses to the specified students based on the course preferences of similar students.

1) *Similarity measure*: The similarity measure between samples is the basis of cluster analysis. When doing classification, a sample is usually considered as 1 vector in an n-dimensional Euclidean space, so the similarity between 2 vectors in n-dimensional Euclidean space can be measured from the following 2 perspectives. One is from the fish degree of vector distance. Second, from the angle of vectors. In particular, since the Euclidean distance in high-dimensional space still satisfies the triangular inequality of distance, the Euclidean distance is the most common method to measure the vector distance in high-dimensional space [6].

a) *Euclidean distance metric*: In high-dimensional space, the Euclidean distance is a measure of the distance between points in vector space that is closest to the intuitive meaning of distance in three-dimensional space. By introducing the concept of Euclidean distance; the vector space has the concepts of length and angle. Suppose the samples x and y are 2 points in an n (n ≥ 1) dimensional Euclidean space.

$$x = [x_1, x_2, x_3, \dots, x_n]^T \quad (1)$$

$$y = [y_1, y_2, y_3, \dots, y_n]^T \quad (2)$$

Then the Euclidean distance is calculated as follows.

$$d(x, y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \dots + (x_n - y_n)^2} = \sqrt{\sum_{i=1}^n (x_i - y_i)^2} \quad (3)$$

b) *Manhattan Distance Metric*: Manhattan distance is also called city block distance assuming that the sample x and y are 2 points in n (n ≥ 1) dimensional space, we get equations (1) and (2), then Manhattan distance is calculated as follows.

$$d(x, y) = |x_1 - y_1| + |x_2 - y_2| + \dots + |x_n - y_n| = \sum_{i=1}^n |x_i - y_i| \quad (4)$$

c) *Chebyshev distance metric*: Chebyshev distance is also an important measure to define the distance of points in

vector space, which takes the maximum value of the distance in the component dimension as Chebyshev distance. Specifically, assuming that the samples x and y are two points in an n (n ≥ 1) dimensional space, equations (1) and (2) are obtained, and the Chebyshev distance is calculated as follows

$$d(x, y) = \lim_{p \rightarrow \infty} (\sum_{i=1}^n |x_i - y_i|^p)^{1/p} \quad (5)$$

d) *Minkowski distance metric*: The Min distance, sometimes referred to as the space-time interval, was first expressed by the Russian-German mathematician H. Minkowski (1864-1909). Assuming that the samples x and y are two points in an n (n ≥ 1) dimensional space, equations (1) and (2) are obtained, and the Minkowski distance is calculated as follows.

$$d(x, y) = (\sum_{i=1}^n |x_i - y_i|^p)^{1/p} \quad (6)$$

From the calculation formula, Min's distance treats the components as obeying the same distribution, and also disregards the difference of the components in the magnitude.

e) *Standardized Euclidean distance Metric*: Similar to the Min distance, the simple Euclidean distance also suffers from the problem of treating the components as obeying the same distribution, and ignores the differences in the components in terms of mean and variance. By first standardizing the components and then calculating the Euclidean distance, an improved standardized Euclidean distance is obtained. Assuming that the samples x and y are 2 points in an n (n ≥ 1) dimensional space, equations (1) and (2) are obtained, and the standardized Euclidean distance is calculated as follows.

$$d(x, y) = \sqrt{\sum_{i=1}^n \left(\frac{x_i - y_i}{s_i} \right)^2} \quad (7)$$

f) *Angle cosine metric*: The angle cosine is a measure of the similarity of sample points from the directional point of view, and is widely used in many fields. Suppose the samples x and y are two points in n (n ≥ 1) dimensional space, and equations (1) and (2) are obtained, then the vector angle cosine is calculated as follows:

$$\cos(\theta) = \frac{\sum_{i=1}^n x_i y_i}{\sqrt{\sum_{i=1}^n x_i^2} \sqrt{\sum_{i=1}^n y_i^2}} \quad (8)$$

From the above equation, the absolute value of the cosine of the vector angle is less than or equal to 1. Its magnitude can reflect the similarity of the two vectors, and the larger the value, the higher the similarity of the two vectors [7]. Moreover, a positive value of the cosine of the vector angle indicates that the two vectors have an isotropic relationship, and vice versa, it indicates that the two vectors. The opposite indicates that the two vectors are negatively related [8]. Considering the similarity measures and the research needs of this paper, the similarity of students is calculated by using the similarity measure based on the cosine of the vector angle.

2) *Student-based collaborative filtering*: In the recommendation system, a sample of 343 senior students and their course selection records from the previous year at Liaoning National Normal College in China was selected as

the basis for this study and the similarity between classmates was calculated by referring to the angle cosine similarity measure. For the purpose of analysis, here is an example of seven students' collaborative over filtering process [9].

Suppose there are seven students, A, B, C, D, E, F and G, who choose the courses they want to take among five courses, a, b, c, d and e. Their course selections are as follows as shown in Fig. 1.

Based on the above students' course selection and the pinch cosine similarity measure, the similarity between students was calculated [10]. For example, the similarity between A and B is shown as below:

$$w_{AB} = \frac{|{a,b,e} \cap {a,d}|}{\sqrt{|{a,b,e}| |{a,d}|}} = \frac{1}{\sqrt{6}} \quad (9)$$

The similarity between A and C is

$$w_{AC} = \frac{|{a,b,e} \cap {c,d}|}{\sqrt{|{a,b,e}| |{c,d}|}} = 0 \quad (10)$$

Similarly, the similarity between A and D is $\frac{2}{\sqrt{6}}$, the similarity between A and E is $\frac{1}{3}$, the similarity between A and F is $\frac{1}{\sqrt{6}}$, and the similarity between A and G is $\frac{2}{3}$

Considering that the above algorithm needs to calculate the similarity between the specified sample (student) and any other sample, in order to improve the computational efficiency of the algorithm, the following improvement scheme is proposed [11].

Step 1: Create a course-to-student reverse lookup table.

Step 2: Build the student's congruence matrix based on the backwards checklist

Based on the selected status of each course in the sample, a backward checklist is created as shown in Fig. 2.

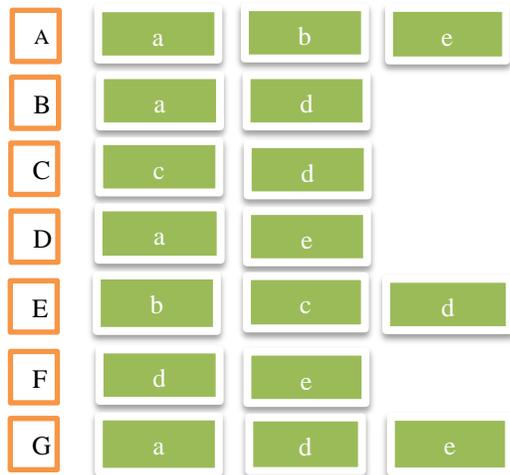


Fig. 1. Sample Student Course Selection List.

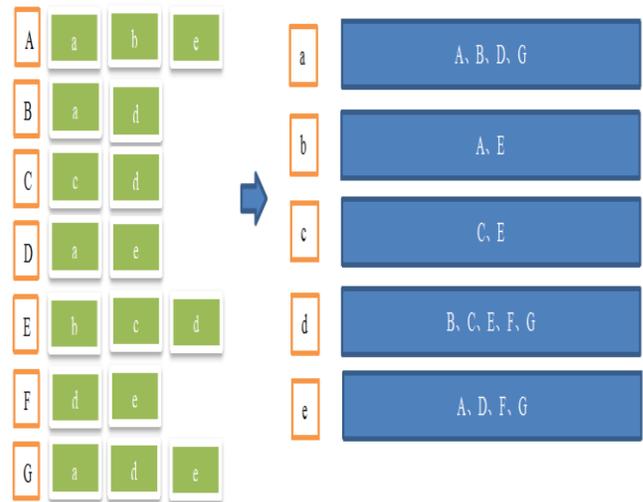


Fig. 2. Sample Student Course Selection Backward Checklist.

Based on the inverted table in Fig. 2, the same present matrix was generated for different students, and the results are shown in Table I.

The co-occurrence table was normalized and the results are shown in Table II.

3) Collaborative course-based filtering: When designing a recommendation system based on collaborative filtering, it is necessary not only to find out the near-neighbor students of a given student based on behavioral data, but also to evaluate the similarity of courses. In evaluating the similarity of courses, the algorithm uses students' ratings of courses to evaluate the similarity between courses [12]. In evaluating course similarity, the algorithm uses students' ratings of courses to evaluate the similarity between courses. To demonstrate the algorithm flow, the analysis is still based on the course selection records of seven students, A, B, C, D, E, F, and G, in five courses, a, b, c, d, and e. The records are analyzed, and their course selections are shown in Fig. 3.

TABLE I. CO-PRESENTATION MATRIX OF THE SAMPLE STUDENTS

	A	B	C	D	E	F	G
A		1		1	1	1	1
B	1		1	1	1	1	1
C		1			1	1	1
D	1	1				1	1
E	1	1					
F	1	1	1	1	1		1
G	1	1	1	1	1	1	

TABLE II. NORMALIZED MATRIX OF THE SAME-PRESENT MATRIX OF THE SAMPLE STUDENTS

	A	B	C	D	E	F	G
A		0.2		0.2	0.2	0.2	0.2
B	0.167		0.167	0.167	0.167	0.167	0.167
C		1			0.25	0.25	0.25
D	0.25	0.25				0.25	0.25
E	0.167	0.167	0.167	0.167		0.167	0.167
FF	0.167	0.167	0.167	0.167	0.167		0.167
GG	0.167	0.167	0.167	0.167	0.167	0.167	

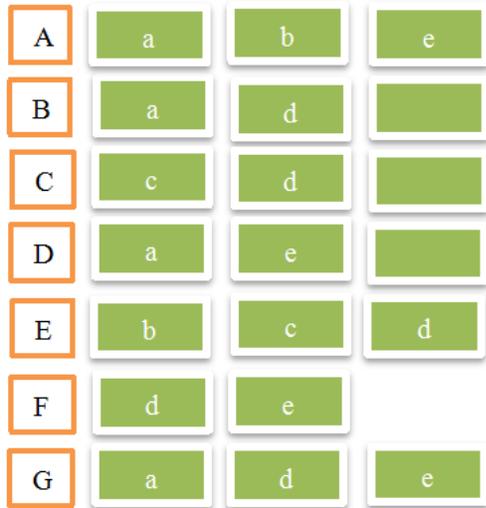


Fig. 3. Sample Student Course Selection Records.

Then, the co-occurrence matrix of the course is normalized. The formula is as follows:

$$w_{ij} = \frac{|N(i \cap j)|}{|N(i)|} \quad (13)$$

The results of the treatment are shown in Table III.

TABLE III. NORMALIZED MATRIX OF COURSE CO-OCCURRENCE MATRIX

	a	b	c	d	e
a		0.33		0.33	0.33
b	0.25		0.25	0.25	0.25
c		0.5		0.5	
d	0.25	0.25	0.25		0.25
e	0.25	0.25		0.25	

After obtaining the students' ratings of the courses, and the similarity between the courses, estimating the preference of students u for the alternative course o [13].

The formula is as follows.

$$p_{u,o} = \sum_{i \in C(0,N)} w_{u,i} p_{u,i} \quad (14)$$

In Equation (14), the measure is the degree of preference of student u for course o . Calculating the degree of preference of a

given student for all alternative courses is the basis for constructing the student course selection recommendation system [14].

In addition, $C(0, N)$ in equation denotes the set of courses similar to the alternative course o , $w_{u,i}$ represents the similarity between the alternative course o and course i , and $p_{u,i}$ is the degree of preference of student u for course i calculated based on the nearest neighbor students [15].

B. Overview of Recommendation Systems

Currently, general education platforms use the course recommendation method based on data statistics, ranking the platform courses based on the number of course selections, and recommending the courses that most platform users are interested in, i.e. the popular course selection list. Among the current recommendation systems, collaborative filtering-based recommendation algorithms are the most widely used [16]. The Recommended methods have their own advantages and disadvantages, in order to better understand their interrelationship and their respective advantages, the two types of algorithms are compared as shown in Table IV.

There are two types of collaborative filtering methods commonly used.

1) user-based collaborative filtering (User-based CF) algorithm, which judges the user's favorite degree of the project through the user's historical behavior, calculates the relationship between users according to different users' preferences for the same project, and recommends the project among users with the same preferences [17].

2) Item-based algorithm focuses on Item, whose similarity is mainly based on its inherent feature values, so it can be classified according to its feature values, calculate the proximity between them, and give suggested results. Since the classification of item is more stable, it can be pushed offline [18].

The comparison of user-based and item-based advantages and disadvantages, and the scope of application have been summarized as shown in Table V.

TABLE IV. COMPARISON OF CLASSIC RECOMMENDATION METHODS

Comparison of Classic Recommendation Methods		
Recommended method	Advantages	Disadvantages
Collaborative filtering recommendations	The ideas are relatively simple and easy to understand; No domain knowledge required; Good comprehension; Can handle unstructured data	Cold start problems; Data sparsity ;
Content-based recommendations	Avoid cold starts and data sparsity;	Difficulty in processing unstructured data Complex feature extraction Recommendation information is prone to over-fitting;

TABLE V. COMPARISON OF USER-BASED AND ITEM-BASED ALGORITHM

Comparison of User-Based And Item-Based Algorithm		
Recommended method	Advantages	Disadvantages
User-based	Push for more socialization. Simply maintain user resemblance Table	Less pronounced specialization. Difficult to provide a push explanation
Item-based	Push specialization. Long-tail item enrichment. Convincing push explanations are available	Unable to adapt to Item updates fast speed

C. The limitation of Previous Studies

The current course selection recommendation system in universities is separate from the online course system. The vast majority of scholars have done a lot of complex work and research in the direction of personalized social network recommendations, e-commerce product recommendations, user emergence mining models, personalized recommendation videos, learning resources, and student user preferences, user profiles, deep recommendation algorithm models, etc.

However, there is very little content related to students' real course selection recommendations, as each school has a different course selection platform. Most schools are concerned about how to complete the task of course selection more easily and quickly, and guarantee the stable operation of the system's course selection platform by randomly drawing lots, or by centrally opening the platform for students to grab courses within a fixed period of time, and do not consider the satisfaction of students' diverse needs through a simple and brutal way.

Algorithms are at the core of personalized recommendation and collaborative filtering techniques. However, how to develop effective and accurate evaluation criteria for the recommendation results of algorithms is an issue that deserves constant attention both for academia and industry. Different evaluation criteria have different focuses, and a single evaluation criterion generally only evaluates a certain aspect of the algorithm, which is more or less deficient. Therefore, how to choose appropriate evaluation metrics to evaluate the recommendation results has a crucial impact on the development of the whole personalized service field [19].

III. METHODOLOGY

Combine course selection function and collaborative filtering and recommendation technology to realize intelligent course selection function. To analyze the trend of students' interest through their information in the system, such as courses taken, grades, ratings, comments, etc., and give them a list of courses. Focusing on the application of pushing process, user-based and item-base are applied to the selection of courses with student and course as the main objects of study. Since the traditional algorithms have some shortcomings, this paper applies item-based weighted and user-based weighted collaborative filtering algorithms to the course selection push system to improve the accuracy. In the design of the course selection push system, consider the major attributes of students,

the relevant attributes of courses and the ratio of students to courses to ensure that the recommendation system can push the set of courses that students are really interested in [20].

A. Data Collection

1) *Selection of data set:* In this paper, we use the data set of the academic system of Liaoning National Normal University to implement the traditional pushing process. A part of the information is extracted as the initial data, and 238 students' evaluation information is obtained. The records were tabulated into students, courses, and ratings tables. Each data file contains the following details.

- Students: Stu-name、Stu-number、Sex、Grade、Score、Zx/Gx (Professional Elective Courses/Public Elective Courses)
- Ratings: Stu-number、Course ID, Rating, and Times.
- Courses: Course-ID、Course-name、Course-category.

The data set is the basis for implementing the course selection push function. Based on the students' course selection and course rating over a period of time, the students' interest level in various elective courses is analyzed and expressed by rank. In this paper, we use a two-dimensional matrix to represent the student's interest in a course, i.e., a $v \times w$ student-course favorite table vw , where v represents the number of students and w represents the number of selected courses. vw value represents the v th student's interest in the w th course. This matrix can be explained by Table VI.

TABLE VI. STUDENT-COURSE GRADING SCALE

	Java	Python	Basketball	H5	Android	PS	Database
student1	5	1	2	1	0	1	4
Student2	2	3	1	2	2	2	5
Student3	2	0	0	4	5	5	2
Student4	3	3	2	3	3	1	3
...
Student N	2	5	0	4	4	4	3

B. Data Description

- Extraction of student attributes and student behavior characteristics:

Define student attributes and behaviors:

student (name、id、sex、grade、Course-name、score、Professional Electives/Public Electives) , For example:Students (Bao Fuyu, 201503, difficult, 21 Computer Applications, Computer Composition, 85, major elective).

- Course Properties

Define course attributes:

Course (course-name, course-id, course-time, score, grade, course-type), such as course (Java programming core technology, F1025, 36 hours 90, 19computer, elective).

C. Research Procedures

The basic framework of intelligent course selection push system is shown in Fig. 4.

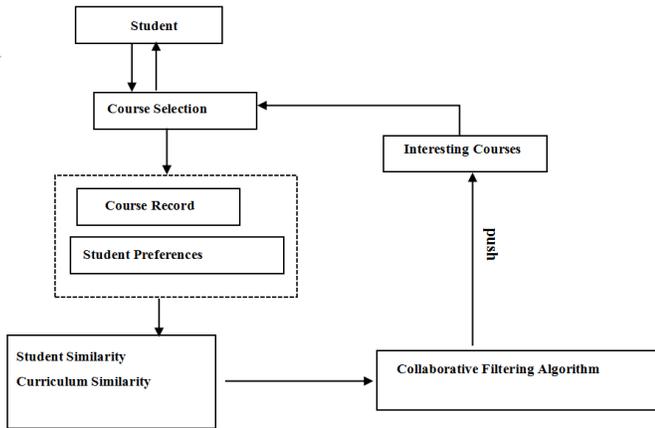


Fig. 4. The Framework of the Push System.

The process of Item-based weighting and User-based weighting push is shown in the following Fig. 5.

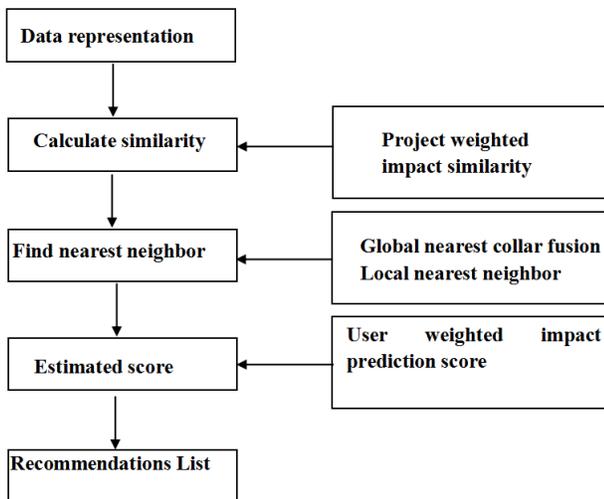


Fig. 5. Item based Weighted and user based Weighted Recommendation Process.

1) *Basic idea:* Consider both course-weighted and student-weighted, and use item-based weighted user cosine similarity and user-based weighted user collaborative filtering to predict the evaluation and improve the accuracy of recommendation [21].

2) *Description of the algorithm process:* Input: target student T, course evaluation form R, course characteristics form A, number of neighbors k;

Output: Top-N recommendation set of target student T;

Step1: Find the set of graded courses of all students, courses and target students T in the system from the course grading table R, and denote them as U_m, I_n, I_T [22].

Step2: The top k students with the highest median value are selected as the closest set of students to student T according to the influence of different course weights to obtain the weighted cosine similarity $N_T = \{j_1, j_2, \dots, j_k\}$.

Step3: For any ungraded course i of the target student T, a weighted approach is used to combine the predicted evaluation of student impact with the predicted evaluation of the student's historical evaluation.

Step4: Select N courses with higher predicted scores as the push result set for target student T.

IV. ANALYSIS AND RESULTS

This chapter is the realization of the system function based on the description and design of the course selection system, then expounds the implementation effect of the recommendation function of the system, and effectively evaluates the recommendation function of the main program of the system according to the evaluation index.

A. Experimental Information

The computer test environment of this machine is shown in Table VII.

TABLE VII. EXPERIMENTAL ENVIRONMENT AND INFORMATION SOURCES

Central processing unit model	Internal Memory (GB)	Operating System	Hard Disk (GB)	Database
Intel(R) Xeon(R) CPU X5650 @ 2.67GHz (12 CPUs)	16GB	Windows11	1TB	Sql Server 2018 R2

B. Data Conversion

The item based weighting and user based weighting recommendation methods are adopted. The item based weighting will affect the proximity value and nearest neighbor selection, and the user based weighting will affect the prediction and evaluation. It is applied to the university educational administration course selection push system to realize the purpose of intelligently pushing elective courses for students. First, get the original data from the database, sort out the original data, retain useful data, delete irrelevant records, and improve the efficiency and accuracy of the algorithm. The data involved in the algorithm studied in this paper are from the educational administration system of Liaoning National Normal University. The data will be processed separately to meet the requirements of the algorithm [23].

From the database, 7 data tables related to students' previous course selection and evaluation data of students' courses are selected. There are 8 tables, from which students' attributes, course selection attributes and students' teaching evaluation information can be obtained. From the interview table collected from students, it can intuitively obtain the students' interest in course in some directions [24]. Obtain the student number, name, course name, major, score, evaluation and other records useful for the algorithm. After re-integrating the records, establish the correlation between tables and rewrite them into the database. See Fig. 6 as follows.

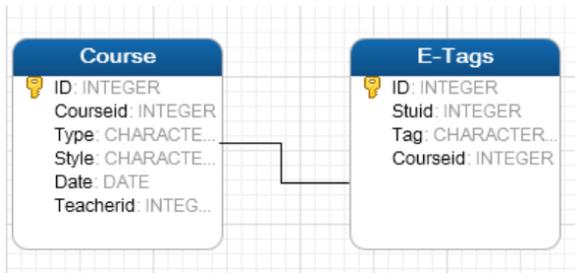


Fig. 6. Association Table after Data Conversion.

C. Accuracy Comparison Results

Experiments show that the improved algorithm can improve the push accuracy. During the experiment, the data comes from.

The data of Liaoning National Normal University educational administration system and the information is collected by questionnaire survey. The test set takes three tenths of the information, the training set takes seven tenths of the information and arbitrarily turns it into five parts, which are expressed as data set 1, data set 2, data set 3, data set 4 and data set 5, respectively. The accuracy flow of push algorithm is shown in Fig. 7.

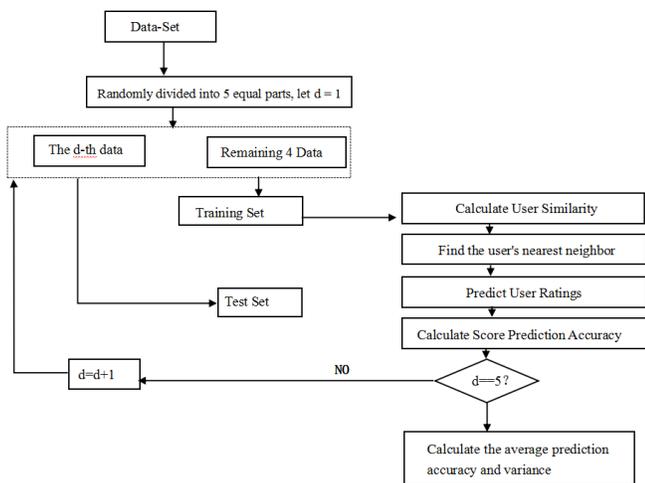


Fig. 7. Accuracy Flow of Collaborative Filtering Algorithm.

V. DISCUSSION

In order to test the prediction accuracy of the recommendation algorithm, the course data set is used for off-line calculation, the student behavior model is established on the training set to predict the student behavior on the test set, and the score prediction accuracy is calculated through the root mean square error RMSE. The detailed experiments are as follows:

- The experimental design uses different data to test the influence of algorithm convergence, and takes the course data set to test, including 343 students and 40 course scores. The data set is divided into two parts: 80% training data and 20% test data.

- Experimental data: under the condition of comprehensive prediction accuracy and efficiency, the experiment carries out three iterative tests of 10K, 100k and 1m data: 5, and 10. The RMSE data is shown in Table VIII.

A. Experimental Analysis

Through the experimental analysis, the following results can be obtained. When the number of iterations increases, the prediction accuracy of RMSE will decrease, and each doubling will decrease by 0.04, indicating that the increase of the amount of data will not significantly reduce the performance of the recommended algorithm, and the increase of the amount of data will make the convergence effect of the algorithm better [25]. At the same time, it can be obtained from the analysis that when the amount of data is the same, the choice of K value affects the prediction accuracy, that is, the smaller the K value, the higher the accuracy, and the larger the K value, the lower the accuracy. Therefore, K is selected as 5 as the number of iterations. The collaborative recommendation algorithm combines the characteristics of User-CF and Item-CF algorithms, and filters the students' information with tags at the initial stage of recommendation, which reduces the search scope to a certain extent. When recommending, filter the initial results of User-CF recommendation, and then make the final recommendation according to the score, so as to make the recommendation effect better, as Fig. 8 and 9.

After many times of verification and improvement, on the whole, the system meets the design requirements. The accuracy of the core algorithm in the recommendation function is evaluated. The experimental results show that the algorithm can produce.

TABLE VIII. RMSE IN DIFFERENT ITERATION DATA

Data-Set	1	2	3	4	5	K
10K	0.9468	0.9411	0.9228	0.9393	0.9265	5
10K	0.9221	0.9254	0.9265	0.9162	0.9232	10
100K	0.8916	0.8870	0.8915	0.8886	0.8891	5
100K	0.8559	0.8671	0.8673	0.8672	0.8655	10
1M	0.8107	0.8382	0.8449	0.8402	0.8322	5
1M	0.8182	0.8182	0.8184	0.8163	0.8153	10

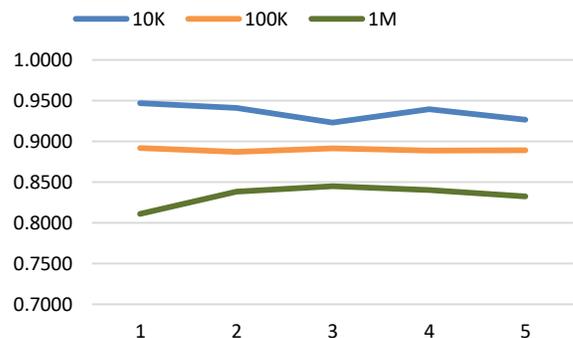


Fig. 8. Association Table after Data Conversion when k=5.

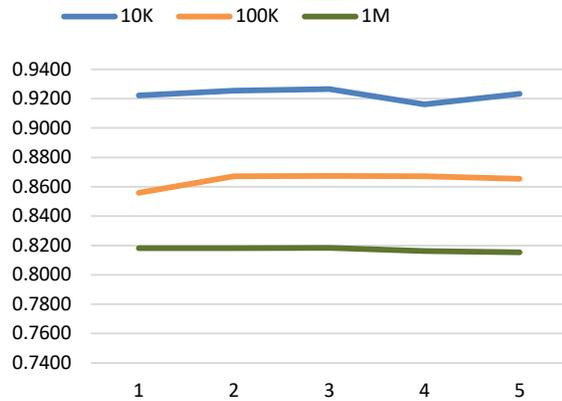


Fig. 9. Association Table after Data Conversion when k=10.

VI. CONCLUSION

In this paper, is used the collaborative filtering algorithm to study students' course selection preferences based on their course selection history data, and build a university course selection recommendation system based on this algorithm. A university course selection recommendation system is built based on this algorithm. The experimental results show that the collaborative filtering algorithm is able to mine and extract the attributes, behavioral characteristics and preferences of students, which can help solve the problems of students' course selection. It can help solve the problem of students' randomness and blindness in course selection, and improve the management efficiency of university education and teaching. In particular, this paper shows that the collaborative filtering algorithm-based student course selection recommendation system can better balance the relationship between student characteristics and course characteristics, and can help achieve the optimal matching between student learning ability and course requirements, student course selection preference and course characteristics, and student ability development and career requirements.

A. Innovation Points

Through the shortcomings of the existing course selection system and the urgent need for an intelligent course selection system, we select the existing course selection system of Liaoning National Normal University and analyze it. Then, introduce collected students' course selection data from Liaoning National Normal University's academic affairs system, pre-processed the data, and compared the user collaborative filtering algorithm based on user weighting with the traditional filtering algorithm through experiments, and the improved way considered the weights of student and course, verified that it can improve the recommendation pushing accuracy in a certain range, and the better pushing result is achieved.

B. Future Work

When looking for nearest neighbors, it can consider the fusion of local nearest neighbors and global nearest neighbors. Global nearest neighbors, local interests are not similar, local nearest neighbors, and global interests are not similar. Through

the optimization algorithm of similarity, the recommendation accuracy can be improved, the data can be used to a greater extent, and the error problem caused by sparse data can be improved. Following problems need to be further studied in the future to supplement and improve the paper: How to make the intelligent course selection recommendation system more accurate and more real-time, and consider the time complexity and space complexity of the algorithm to make the efficiency better?

REFERENCES

- [1] J. Xiao, M. Wang, B. Jiang, and J. Li, "A personalized recommendation system with combinational algorithm for online learning," *Journal of Ambient Intelligence and Humanized Computing*, vol. 9, no. 3, pp. 667–677, 2018.
- [2] R. N. Behera and S. Dash, "A particle swarm optimization based hybrid recommendation system," *International Journal of Knowledge Discovery in Bioinformatics*, vol. 6, no. 2, pp. 1–10, 2016.
- [3] M. K. Najafabadi, A. Mohamed, and C. W. Onn, "An impact of time and item influencer in collaborative filtering recommendations using graph-based model," *Information Processing and Management*, vol. 56, no. 3, pp. 526–540, 2019.
- [4] G. Geetha, M. Safa, C. Fancy, and D. Saranya, "A hybrid approach using collaborative filtering and content based filtering for recommender system," *Journal of Physics: Conference Series*, vol. 1000, pp. 012101–012123, 2018.
- [5] L. Guo, J. LiangYing, Z.Y. Luo, and L. Sun, "Collaborative filtering recommendation based on trust and emotion," *Journal of Intelligent Information Systems*, pp. 1–23, 2018.
- [6] H. Zarzour, F. M. Mohamed, S. Chaouki, and C. Chemam, "An improved collaborative filtering recommendation algorithm for big data," *Computational Intelligence and Its Applications*, pp. 660–668, 2018.
- [7] H. Zhang, T. Huang, Z. Lv, S. Liu, and Z. Zhou, "MCRS: a course recommendation system for MOOCs," *Multimedia Tools and Applications*, vol. 77, no. 6, pp. 7051–7069, 2018.
- [8] R. Logesh, V. Subramaniaswamy, D. Malathi, N. Sivaramakrishnan, and V. Vijayakumar, "Enhancing recommendation stability of collaborative filtering recommender system through bio-inspired clustering ensemble method," *Neural Computing and Applications*, vol. 32, no. 7, pp. 2141–2164, 2020.
- [9] Y. Yang, Y. Xu, E. Wang et al., "Improving existing collaborative filtering recommendations via serendipity-based algorithm," *IEEE Transactions on Multimedia*, vol. 20, no. 7, pp. 1888–1900, 2017.
- [10] H. Li, H. Li, S. Zhang, Z. Zhong, and J. Cheng, "Intelligent learning system based on personalized recommendation technology," *Neural Computing and Applications*, vol. 31, no. 9, pp. 4455–4462, 2019.
- [11] A. Klasnja-Milicevic, M. Ivanovic, B. Vesin et al., "Enhancing e-learning systems with personalized recommendation based on collaborative tagging techniques," *Applied Intelligence*, vol. 48, no. 6, pp. 1519–1535, 2018.
- [12] J. Chen, C. Zhao, L. Ulji, and L. Chen, "Collaborative filtering recommendation algorithm based on user correlation and evolutionary clustering," *Complex and Intelligent Systems*, vol. 6, no. 1, pp. 147–156, 2020.
- [13] L. Jiang, Y. Cheng, L. Yang, J. Li, H. Yan, and X. Wang, "A trust-based collaborative filtering algorithm for E-commerce recommendation system," *Journal of Ambient Intelligence and Humanized Computing*, vol. 10, no. 8, pp. 3023–3034, 2019.
- [14] X. Liu, "A collaborative filtering recommendation algorithm based on the influence sets of e-learning group's behavior," *Cluster Computing*, vol. 22, no. 2, pp. 2823–2833, 2019.
- [15] J. Chen, C. Zhao, L. Ulji, and L. Chen, "Collaborative filtering recommendation algorithm based on user correlation and evolutionary clustering," *Complex and Intelligent Systems*, vol. 6, no. 1, pp. 147–156, 2020.

- [16] Karalis, T., & Raikou, N. (2020). Teaching at the times of COVID19: Inferences and implications for higher education pedagogy. *International Journal of Academic Research in Business and Social Sciences*, 10(5), pp 479-493.
- [17] Wang, X., Zhu, Z., Yu, J., Zhu, R., Li, D., Guo, Q. (2019). A learning resource recommendation algorithm based on online learning sequential behavior. *International Journal of Wavelets, Multiresolution and Information Processing*, 17(2).
- [18] Almusharraf, N., Khahro, S. (2020). Students Satisfaction with Online Learning Experiences during the COVID-19 Pandemic, *International Journal of Emerging Technologies in Learning*, 15(21):pp 246-267.
- [19] Diao, X., Zeng, Q., Duan, H., Song, Z., Zhou, C. (2021). Personalized learning resource recommendation based on course ontology and cognitive ability. *Journal of Computers*, 32(2): pp149-163.
- [20] Liu, Z.Y., Lomovtseva, N., Korobeynikova, E. (2020). Online Learning Platforms: Reconstructing Modern Higher Education, *International Journal of Emerging Technologies in Learning*, 15(13): 4-21.
- [21] Chung, E., Subramaniam, G., & Dass, L. C. (2020). Online Learning Readiness among University Students in Malaysia amidst COVID-19. *Asian Journal of University Education*, 16(2), pp 46-58.
- [22] Bao, W. (2020). COVID-19 and online teaching in higher education: A case study of Peking University. *Human Behavior and Emerging Technologies*, 2(2), pp113-115.
- [23] Adnan, M., & Anwar, K. (2020). Online Learning amid the COVID-19 Pandemic: Students' Perspectives. *Online Submission*, 2(1), 45-51.
- [24] Sahbaz, A. (2020). Views and evaluations of university students about distance education during the COVID-19 pandemic. *Educational Process: International Journal (EDUPIJ)*, 9(3), pp184-198.
- [25] Al-Salman, S., & Haider, A. S. (2021). Jordanian University Students' Views on Emergency Online Learning during COVID-19. *Online Learning*, 25(1), pp286-302.

Intelligent Interfaces for Assisting Blind People using Object Recognition Methods

Jamil Abedalrahim Jamil Alsyayadeh^{1*}
Hasvini Baskaran⁴

Department of Electronics & Computer Engineering
Technology, Fakulti Teknologi Kejuruteraan Elektrik &
Elektronik (FTKKE), Universiti Teknikal Malaysia Melaka
(UTeM), Melaka, Malaysia

Irianto²

Department-General Education, Faculty of Resilience,
Rabdan Academy, Abu Dhabi, United Arab Emirates

Maslan Zainon³

Department of Electrical Engineering Technology
Fakulti Teknologi Kejuruteraan Elektrik & Elektronik
(FTKKE), Universiti Teknikal Malaysia Melaka (UTeM)
Melaka, Malaysia

Safarudin Gazali Herawan⁵

Industrial Engineering Department
Faculty of Engineering, Bina Nusantara University
Jakarta, Indonesia 11480

Abstract—Object recognition method is a computer vision technique for identifying objects in images. The main purpose of this system build is to put an end to blindness by constructing automated hardware with Raspberry Pi that enables a visually impaired person to detect objects or persons in front of them instantly, and inform what is in front of them through audio. Raspberry Pi receives data from a camera then processes it. In addition, the blind will listen to a voice narration via an audio receiver. This paper's key objective is to provide the blind with cost-effective smart assistance to explore and sense the world independently. The second objective is to provide a convenient portable device allows users to recognise objects without touch, having the system determine the object in front of them. The camera module attached in Raspberry Pi will capture image and the processor will then process it. Subsequently, the processed image sends data to the audio receiver narrating the detected object(s). This system will be very useful for a blind person to explore the world by listening to the voice narration. The generated voice narration after processing the image will help the blind to visualise objects in front of them.

Keywords—Object recognition method; computer vision; blind people; image processing; Raspberry Pi; Pi camera; smart assistance; portable device; voice narration; visualise

I. INTRODUCTION

Physical movement is a challenge for the blind. Visual disability stands out from the many extreme obstacles affecting a person. This is clearly explained in this research study about the visually impaired patients. They face physical and social constraints accrued from their visual loss, and they need to improve on their health and independence [1][2][3]. Designing a gadget to help the blind is not something new. Various technologies exist to help the visually impaired, and as innovations increasingly propelled, ideas appear to provide intriguing measures to help the visually impaired. In any case, designing a device to aid blindness comes with a price and does not come easy, as it is often regarded as a so-called luxurious item in most developing nations.

As indicated by the World Health Organization (WHO) research study, [4][5] it has been estimated that over 1.3 billion people around the world have some form of vision impairment. Roughly, 80% of all kinds of vision debilitation are viewed as avoidable. Additionally, the WHO also mentioned that most visually impaired adolescents would require visual recovery intercessions for self-improvement. Nevertheless, it is most often that visual recovery treatments come with substantial hospital expenses, and with 90% of disabled people living with financial difficulties, visual restoration is not the best alternative for all. For full psychological improvement and better independence without bearing costly bills, blind people require an assistive device that helps them with their daily activities. There exist many assistive devices for visually impaired people and it became the inspiration in the background of this research study. Smart assistance such as a smart and autonomous walking stick, smart glasses/spectacles, or prosthetics [6][7][8][9]. The assistance from another individual is not always accessible, and are unfavoured by visually impaired individuals that search for freedom, without having to bear the cost of such expensive smart assistance equipment as well. We propose an audio receiver for blind people that uses real-time smart assistance interfaces and object recognition technologies as a solution to this occurring issue. Our system mainly consists of two components, Raspberry Pi and Pi camera. The smart assistance audio guidance was developed to assist users to determine objects in front of them and help them visualise the environment around them. The camera in the processor will capture image, then processes that image, and a voice narration will be sent through audio receiver.

The main objective is to provide the blind cost-effective smart assistance to explore and feel the world independently. This enables the blind to visualise their surroundings and afford current technologies. Additionally, we also aim to provide a portable device which is easy to utilise and permits them to recognise objects without touching, and describes the surroundings in front of them.

*Corresponding Author.

This system is used to assist blind people with voice narration, processed by the Raspberry Pi processor. This portable electronic device's purpose is to give voice narration informing what is in front of them. An important objective is to provide a portable device that is simple to use and low cost and affordable smart assistance to blind people. Another goal is to extend the computerised electronic travel aid for the blind by applying real-time object recognition technology. This blind guidance system is solid and financially perceptible. Real-time based smart assistance interfaces the audio receiver for blind people with voice narration by using object recognition methods to provide the blind with cost-effective smart assistance to explore and sense the world independently. Audio guidance helps them to know what is happening around them and it helps them to visualise their surroundings. By using real-time, the system will recognise the objects faster.

The remaining of this paper has been organized as follows: Section 2 discusses the related works. The background of the study is described in Section 3. Section 4 described the system implementation and testing. Section 5 described the results and discussion and finally, the conclusion is described in Section 6.

II. RELATED WORK

There are a lot of assistive devices for visually impaired people to sense the world independently. All these devices rely mainly on ultrasonic sensors and Brailleing.

A. EyeCane and EyeMusic

Maidenbaum et al. [10] designed EyeCane and EyeMusic to improve upon, or likely be within the far distant future, to update the traditional white cane. By applying statistics at visually far distances (5 meters) and greater angles, and most significantly by means of discarding contacts among the cane, and the user's surroundings in cluttered or indoor environments. The EyeCane converts point-distance information into aural and tactile signals. The Prototype of EyeCane and EyeMusic is shown in Fig. 1. The tool can provide distance information to the customer from two different directions at the same time: immediately in advance for long-distance perception and detection of waist-height obstacles, and pointing downward at a 45° angle for ground-level evaluation.



Fig. 1. Prototype of EyeCane and EyeMusic.

B. Blitab

Blitab is a device nicknamed "the iPad for the visually impaired". It appears similar to a digital book, however, its screen utilises smart liquids that protrude tactile pixels to show braille letters, making it conceivable for the blind to see entire pages of braille message at once. Perkins-style keyboard application, text-to-speech yield, and touch navigation provide a completely new user experience for braille and non-braille blind individuals. It empowers the fast conversion of any content into braille. Blitab is a platform for all current and future programming applications for visually impaired people, it is not only a tablet. The Prototype of Blitab is shown in Fig. 2.

Blitab is the world's first real tactile tablet designed specifically for the blind and visually impaired. The device's revolutionary smart liquid technology also allows it to display material images for blind people who do not use braille [11].

C. BrainPort V100

According to Grant et al. [12], BrainPort V100 is an oral electronic vision aid that uses electro-tactile stimulation to help profoundly blind people with direction, mobility, and object recognition. The device is used in conjunction with other assistive devices like a normal white cane or a guide dog.

It deciphers digital data from a wearable camcorder into delicate electrical incitement designs on the outside of the tongue. Users feel moving bubble-like patterns on their tongue then they figure out how to interpret or visualise according to the shape, size, area, and movement of articles in their condition. A few clients have portrayed it as having the option to "see with your tongue". What makes it extraordinary is seeing with your mouth may appear to be outlandish at first, yet with at least 10 hours of one-on-one instructional courses, wearers can figure out how to comprehend the shivers and "see" where objects are found, yet additionally, their size, shape and in the event that they are moving. In a clinical preliminary, 69% of members had the option that effectively recognises protests in an acknowledgment test following one year of preparing with the BrainPort. The Prototype of BrainPort V100 is shown in Fig. 3.



Fig. 2. Prototype of Blitab.



Fig. 3. Prototype of BrainPort V100.

TABLE I. COMPARISON BETWEEN EXISTING SYSTEM

System	Devices needed	Cost	accessibility	purpose
BrainPort V100	Headset, Intra Oral Device(IOD)	Expensive (\$10000)	Controller	To provide oral electronic vision aid
EyeCane& EyeMusic	Infrared emitters, web camera, smartphone, headset	Low cost	Infrared sensors	To provide navigation control and identifies colour, shape and location of objects.
Blitab	Touch screen tablet, Braille lines	Low cost(\$500)	Braille display	Displays tactile images to blind people.
Proposed method	Raspberry Pi 3, Pi camera module, Headset	Cheap (\$100)	Text-to-Speech module	The generated narration will be the final output that is transmitted to the user through a headset.

The difference between existing systems is shown in Table I.

III. BACKGROUND OF THE STUDY

Object Recognition is a method used in image processing to recognise real objects. This method is clearly explained in a

research study about the importance of process that will help blind people to identify their daily items that are commonly used. Our system provides some kind of visual aid that recognises objects dynamically [13]. The algorithm used in this system analyses the object. For instance, a blind person is sitting on his dining table. He has multiple objects in front of him such as bottle, chair, dining table, etc. Therefore, our system will help him by narrating what is in front of them. Text-to-Speech module is used to convert text to speech. The text that is written in text file is the output of object detection. Google API is used for conversion of Text-to-Speech dynamically, provided that the internet connection is stable. This has been studied from a research that explains about Google API that is used for text-to-speech [14]. For example, if the camera captures a book in front of it, it detects the book and converts it into text from the image captured. The text will be written in a text file and then converted to speech by using Google Text-to-Speech. The architecture of this proposed system is the Raspberry Pi board. Raspberry Pi controller controls the system and activates the output and sends the instructions. The detailed specifications of Raspberry Pi 3 B+ consists of: four USB ports, an Ethernet port, forty GPIO pins, SD card slot, SOC (system on a chip), a DSI display interface, HDMI port, LAN controller, audio jack, CSI camera interface, RCA video socket, and 5V micro USB connector [15].

The Block diagram of the object recognition process is shown in Fig. 4.

The Pi camera is connected to a CSI camera interface of Raspberry Pi processor. The processor has an operating system named Raspbian, which process the image, voice narration and other conversions. The headset will connect to an audio jack for audio output. Once the system components activate, the camera module will begin a video stream of its front view, and the image in video will be processed. Before this process starts, the Raspberry Pi will create a video frame, activates "cv" environment, and runs the python script to activate the system. Thereafter, the processed image undergoes object detection for image classification and recognition. Hence, the image in the video will detect through real-time object recognition, and the label of each object will be printed in a text file, which is used for voice narration. The labels in the text file use Google Text-to-Speech for voice narration. The generated narration will be the final output that is transmitted to the user through a headset.

The flowchart of the object recognition process is shown in Fig. 5.

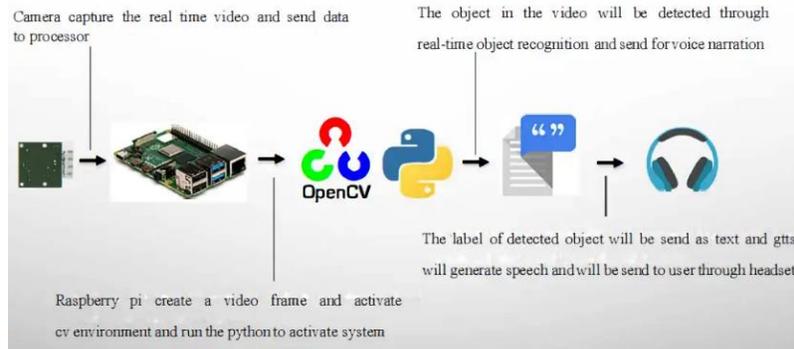


Fig. 4. Block Diagram of the Object Rprocess.

IV. SYSTEM IMPLEMENTATION AND TESTING

A. Hardware Implementation

The necessary components in developing this system consist of a Raspberry Pi and Pi camera. The New Out of Box Software (Noobs) is installed on an SD card to format the Raspberry Pi that will be fixed in the Raspberry Pi, as studied in the manual that was given to study about Raspberry Pi startup [16]. Noobs contain Java SE Platform Products. It is an operating system installer with Raspbian pre-loaded. Once done, this Raspberry Pi will connect to power, start to boot, and be ready to use the operating system, whereas the Pi camera will be configured beforehand. The camera's interfacing option in Raspbian OS will be enabled manually to allow the camera to work with the system. Once the configuration is done, the Raspbian enables the camera. The image captured after configuring the Pi camera is shown in Fig. 6.

B. Software Implementation

The Raspbian operating system is used in the Raspberry Pi 3 model B+ as a platform to run this system; which is, the platform to create, run, and troubleshoot the coding of the software that has been used. Python IDLE software was used to build this system. Python IDLE ran in an OpenCV environment. OpenCV was created to provide a common infrastructure for computer vision applications such as deep learning, optical character recognition (OCR) and object detection, and more as explained in the article [17][18]. OpenCV-Python is the Python API for OpenCV. It's a Python bindings library aimed in solving computer vision challenges. Python has been enhanced with C/C++, enabling programmers to write/express code and develop Python wrappers that can be used as Python modules, as stated in the article named Python. It is packaged as an optional part of the Python packaging with many Linux distributions [19][20]. The actual code will run in the background of the CV environment. To write the necessary codes to run the system, the Python 3.7.3 was used. To capture the image, a Pi camera connected to a Raspberry Pi was used. Furthermore, code will be used to initialise the captured image.

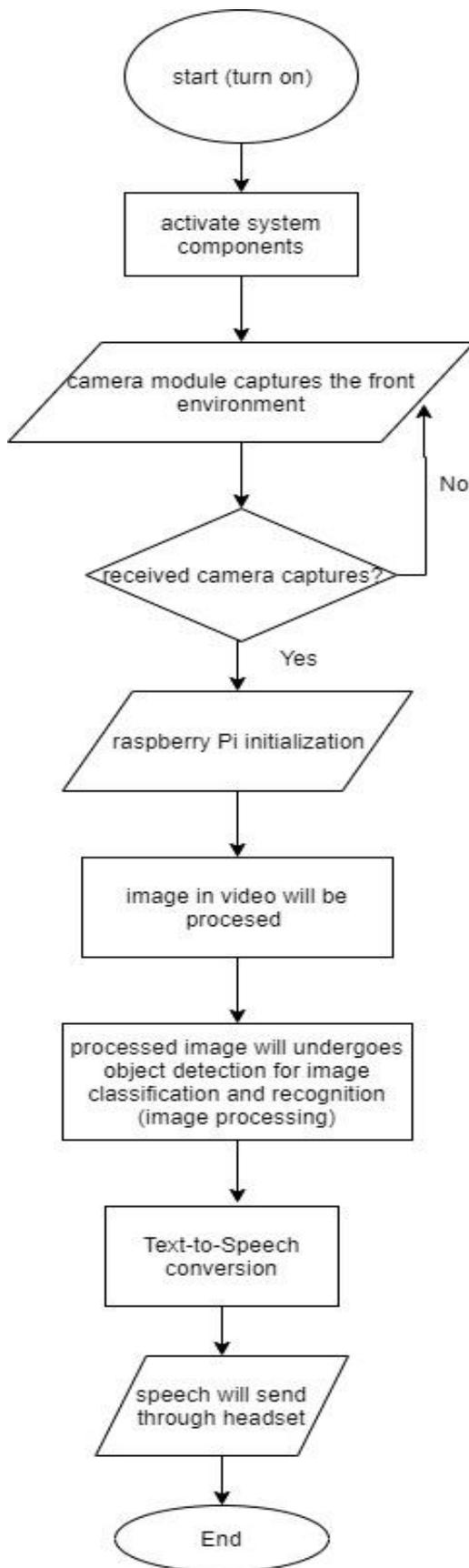


Fig. 5. Flowchart of the Object Recognition Process.



Fig. 6. Image Captured after Configuring the Pi Camera.

V. RESULT AND DISCUSSION

Here is a sample of coding and result for this proposed system. Some discussions are added up as an explanation to understand its function clearly.

A. Coding

Partially applied programming codes are displayed below in Fig. 7 to Fig. 10.

```
ap = argparse.ArgumentParser()
ap.add_argument("-p", "--prototxt", required=True,
                help="path to Caffe 'deploy' prototxt file")
ap.add_argument("-m", "--model", required=True,
                help="path to Caffe pre-trained model")
ap.add_argument("-c", "--confidence", type=float, default=0.2,
                help="minimum probability to filter weak detections")
args = vars(ap.parse_args())
```

Fig. 7. Construction of the Argument Parse.

The preceding code demonstrates how to construct an argument parse to parse the arguments. 'ArgumentParser()' converts the argument value from a string into some other type. The first line sets up an argument parser, followed by three mandatory command-line arguments. Firstly, the 'prototxt' is the path to the Caffe prototxt file which is known as the solver.prototxt, secondly, a configuration file, whereas '— model' is the path to the pre-trained model and thirdly, the '— confidence' is minimum probability threshold when filtering weak detections and it is set to 20% by default.

```
CLASSES = ["background", "aeroplane", "bicycle", "bird", "boat",
           "bottle", "bus", "car", "cat", "chair", "cow", "diningtable",
           "dog", "horse", "motorbike", "person", "pottedplant", "sheep",
           "sofa", "train", "tvmonitor"]
COLORS = np.random.uniform(0, 255, size=(len(CLASSES), 3))

# load our serialized model from disk
print("[INFO] loading model...")
net = cv2.dnn.readNetFromCaffe(args["prototxt"], args["model"])
```

Fig. 8. Initialisation of 'Classes'.

These lines of code initialise 'CLASSES', class labels, and equivalent COLORS, for on-frame text and bounding boxes. Furthermore, the last line loads the serialised neural network model.

```
for i in np.arange(0, detections.shape[2]):
    # extract the confidence (i.e., probability) associated with
    # the prediction
    confidence = detections[0, 0, i, 2]

    # filter out weak detections by ensuring the 'confidence' is
    # greater than the minimum confidence
    if confidence > args["confidence"]:
        # extract the index of the class label from the
        # 'detections', then compute the (x, y)-coordinates of
        # the bounding box for the object
        idx = int(detections[0, 0, i, 1])
        box = detections[0, 0, i, 3:7] * np.array([w, h, w, h])
        (startX, startY, endX, endY) = box.astype("int")

        # draw the prediction on the frame
        label = "{}: {:.2f}%".format(CLASSES[idx],
                                   confidence * 100)
        cv2.rectangle(frame, (startX, startY), (endX, endY),
                    COLORS[idx], 2)
        y = startY - 15 if startY - 15 > 15 else startY + 15
        cv2.putText(frame, label, (startX, y),
                    cv2.FONT_HERSHEY_SIMPLEX, 0.5, COLORS[idx], 2)
```

Fig. 9. Looping the Detection.

This part explains how this system is able to detect numerous objects in a single image. First step is to loop over the detections. The chance of each detection will be checked and tallied with confidence. If the confidence exceeds the

threshold, the prediction will be displayed in terminal and drawn on the frame. Detections will undergo loops and its confidence value is extracted in each loop. Therefore, the class label index is extracted if the confidence level is greater than the minimal threshold, as well as the bounding box coordinates surrounding the detected objects that have been computed too, and a rectangle displaying text is created on the detected object. Labels containing CLASS name and confidence build, and displayed as the processed-colored rectangle created around the object. Finally, the system computes the colored text that was generated onto the frame by using the y-value.

```
from gtts import gTTS
from time import sleep
import os
import pygame

file = open("label.txt", "r").read().replace("\n", " ")
language = 'en'
tts = gTTS(text = 'hi there. i am here to assist you to look the world in front of you, there is' + str(file) + 'in front of you', lang = 'en', slow = False)
tts.save('voice.mp3')
os.system("start /home/pi/Desktop/real-time-object-detection/voice.mp3")
print(file)
```

Fig. 10. Voice Narration.

To enable the Raspberry PI to "talk", the Google Text to Speech (gTTS) module is used in Python is used and also imported into the Raspbian system. This is used to command the system to read the image classification result that has been written after the real time object detection process. To put it simply, this python coding aims to read the text file and then create a voice narration.

B. Result

The system has been tested and its functionality has been demonstrated as per the design. The system has been able to operate as designed, thanks to the combination of software and hardware components. The system interface with Pi camera will capture the front environment and the data will be transferred to the processor to process the image. Object recognition methods will enable image processing, convert it to text, and use Google-Text-to-Speech to create the voice narration and send it to the user through an audio receiver.

The detected objects in frame are shown in Fig. 11.

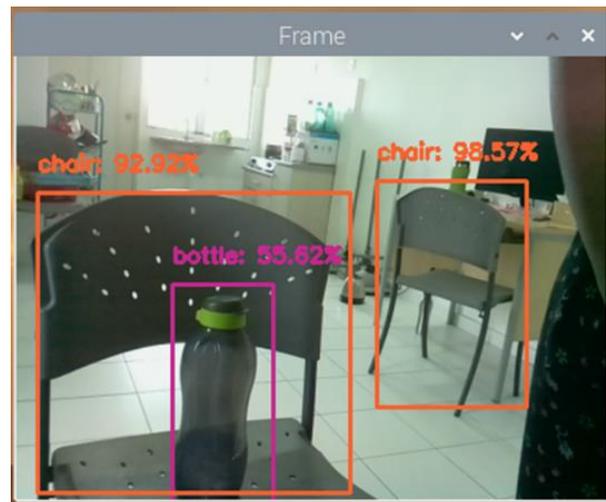


Fig. 11. Shows Objects Detected in Frame and its Confidence Percentage Value of the Detected Objects with its Label of Classes.

```
[INFO] loading model...
[INFO] starting video stream...
table: 57.87%
bottle: 26.65%
person: 99.65%
person: 98.43%
tvmonitor: 99.80%
tvmonitor: 99.46%
chair: 28.39%
person: 96.32%
chair: 54.55%
table: 33.28%
sofa: 51.36%
person: 78.48%
person: 99.10%
person: 97.11%
person: 99.15%
table: 26.27%
person: 49.29%
person: 95.13%
person: 88.81%
```

Fig. 12. Shows Label Classes will be Printed along with the Detection Confidence Percentage Value.

From Fig. 12, the results from the system are collected and the percentage of confidence value obtained, to show the objects detected along with its confidence percentage value. Confidence value is the probability that a bounding box containing an object and it is predicted by a classifier. The object in the bounding box would return many predictions, but out of those, most of them will have a very low confidence value associated. Hence, only predictions above 20% confidence is reported, as fixed in the python coding itself. That is how the object detection algorithm returns values after confidence thresholding, once the video stream starts in our system. As previously indicated in Fig. 11, the objects in the bounding box are correct, due to the quantifying the predictions.

To confirm the predictions, the correctness value of each object detection should be obtained. The measurement that determines the correctness of the bounding box is the Intersection over Union (IoU). IoU [21][22][23] is the ratio between the intersection, and the union of speculated boxes and ground truth boxes. The IoU's calculation is shown below in Fig. 13.

Consequently, the correct detections will be identified, and then its precision and recall will be calculated. To calculate precision and recall, the True Negatives, False Negatives, True Positives and False Positives will be identified. To obtain True Positives and False Positives, IoU will be used and the detection will be identified to determine whether it is correct (True Positive) or not (False Positive). The used threshold is 0.2, if IoU is > 0.2, it is considered a True Positive. Else, it will be considered as a False Positive. The COCO (Common Objects in Context) evaluation metric suggests measurements through various IoU thresholds.

To calculate the recall, the count of Negatives is required because not every part of the image in the video stream frame detected is an accepted object or is considered a negative. False Negatives will only be measured if the objects detected by our system are missed out. The recall is calculated as the ratio between the number of correct predictions (A) (True Positive) and the missed detections (False Negatives). The correct predictions for each class in the video stream will be commutated after calculation of IoU using the ground truth boxes for each positive detection box that the system has reported. So, with this, the IoU threshold (0.2). Therefore, the formulas will be as below.

$$Precision = \frac{TP}{(FP+TP)} \tag{1}$$

$$Recall = \frac{TP}{(TP+FN)} \tag{2}$$

Subsequently, the Mean Average Precision (mAP) is calculated in Table II. mAP is used in the domains; Information Retrieval and Object Detection. These two domains have separate ways to calculate mean average precision. Object detection of mAP is formalised in the PASCAL Visual Object Classes (VOC). PASCAL VOC provides a common dataset of images and annotations, as well as a standard evaluation to the vision and machine learning communities [24][25]. The average precision for all object types is shown in the table below. The PASCAL VOC dataset's mAP was found to be 0.665. The best mAP value at the moment is reported to be 0.739.

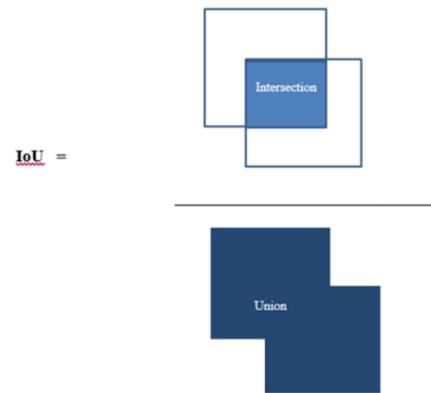


Fig. 13. IoU Calculation.

TABLE II. AVERAGE PRECISION FOR ALL CLASSES

No.	Class	Average Precision
1	Train	0.542
2	Bicycle	0.636
3	Dog	0.818
4	Diningtable	0.534
5	Aeroplane	0.727
6	Chair	0.909
7	Person	0.909
8	Tvmonitor	0.633
9	Bus	0.726
10	Sofa	0.710
11	Bird	0.727
12	Cow	0.632
13	Bottle	0.909
14	Pottedplant	0.359
15	Boat	0.544
16	Car	0.634
17	Cat	0.272
18	Sheep	0.633
19	Motorbike	0.724
20	Horse	0.726

VI. CONCLUSION

The system's goal of providing intelligent help for visually impaired people through real-time based object recognition has been successfully developed. Most of the important details in the general theory of design and execution have also been introduced throughout this article. From the theory to the practical realisation of this category of smart assistance for visually impaired people, these developments involve a variety of technical and coding details. From the testing and result analysis, the designed system's functionality is advanced and helps the visually impaired people to know what is in front of them. According to the data analysis based on Table II, the average precision for all classes is shown. Occasionally, detecting precision is not as precise as it should be, because the object is detected using values assigned by the system. Additional objects of comparable size or shape may also be detected with incorrect predictions. The strength of this system is users are able to listen to the voice narration audio that informs them what objects are in front of them. The Mean Average Precision (mAP) was calculated. The PASCAL VOC dataset's mAP was found to be 0.665. The best mAP value at the moment is reported to be 0.739.

The limitation of this system is it only has one pi camera interfaced to raspberry pi to capture video stream. So the scope only for blind people since the system included with pre-trained model that used for object detection. The most important recommendation for improvement, is about a future work development by implementing the Non-Maximum Suppression, making the regions more accurate. The object detection algorithm is good but not very accurate sometimes, because the regions reduce the ratio of algorithm. Furthermore, the development of this system should include a more pre-trained model in larger numbers. Lastly, the system can be improved by being cloud based, that way, all the data that had been captured will be saved in the cloud, and it will be easy for the user's guardian to acknowledge the details this system has generated, and can include localisation to know the location of the user travelled.

ACKNOWLEDGMENT

The authors would like to thank Centre for Research and Innovation Management (CRIM) for the support given to this research by Universiti Teknikal Malaysia Melaka (UTeM). We thank also those who contributed in any other forms for providing their continuous support throughout this work.

REFERENCES

- [1] M. Glatz, R. Riedl, W. Glatz, M. Schneider, A. Wedrich, M. Bolz, R. W. Strauss. "Blindness and visual impairment in Central Europe", *PLoS one*. 2022; 17(1):e0261897. <https://doi.org/10.1371/journal.pone.0261897>.
- [2] SR. Flaxman, RRA. Bourne, S. Resnikoff, P. Ackland, T. Braithwaite, MV. Cicinelli, et al. Global causes of blindness and distance vision impairment 1990–2020: a systematic review and meta-analysis. *Lancet Glob Health*. 2017;5(12):e1221–e34. pmid:29032195.
- [3] TY. Wong, J. Sun, R. Kawasaki, P. Ruamviboonsuk, N. Gupta, VC. Lansingh, et al. Guidelines on Diabetic Eye Care: The International Council of Ophthalmology Recommendations for Screening, Follow-up, Referral, and Treatment Based on Resource Settings. *Ophthalmology*. 2018;125(10):1608–22. pmid:29776671.
- [4] B. Thylefors, A. D. Negrel, R. Pararajasegaram, and K. Y. Dadzie. "Global data on blindness", *Bulletin of the World Health Organization*. vol. 73, no. 1, pp. 115–121, 1995. PMID: 7704921; PMCID: PMC2486591.
- [5] A. Hydera, I. Mactaggart, S. J. Bell, J. A. Okoh, S. I. Olaniyan, M. Aleser, H. Bobat, A. Cassels-Brown, B. Kirkpatrick, M. J. Kim, I. McCormick, H. Faal, M. J. Burton. "Prevalence of blindness and distance vision impairment in the Gambia across three decades of eye health programming". *The British Journal of Ophthalmology*. 2021 Dec;bjophthalmol-2021-320008. DOI: 10.1136/bjophthalmol-2021-320008. PMID: 34949578.
- [6] M. A. Ikbali, F. Rahman, M. R. Ali, M. H. Kabir and H. Furukawa, "Smart walking stick for blind people: An application of 3D printer", *Proc. SPIE 10167 Nanosensors Biosensors Info-Tech Sensors 3D Syst.*, pp. 101670T, Apr. 2017.
- [7] A. Krishnan, G. Deepakraj, N. Nishanth, and K. M. Anandkumar, "Autonomous walking stick for the blind using echolocation and image processing." 2016. 2016 2nd International Conference on Contemporary Computing and Informatics (IC3I), pp. 13-16.
- [8] J. Bai, S. Lian, Z. Liu, K. Wang, and Di. Liu, "Smart guiding glasses for visually impaired people in indoor environment," *IEEE Transactions on Consumer Electronics*. vol. 63, no.3, pp. 258–266, 2017.
- [9] Z. O. Abu-Faraj, E. Jabbour, P. Ibrahim, and A. Ghaoui, "Design and development of a prototype rehabilitative shoes and spectacles for the blind," 2012. 5th International Conference on Biomedical Engineering and Informatics, BMEI, pp. 795-799.
- [10] S. Maidenbaum, S. Hanassy, S. Abboud, G. Buchs, D. R. Chebat, S. Levy-Tzedek, A. Amedi. "The "EyeCane", a new electronic travel aid for the blind: Technology, behavior & swift learning," *Restorative neurology and neuroscience*. vol. 32, no. 6, pp. 813–824, 2014.
- [11] J. L. Robinson, V. Braimah Avery, R. Chun, G. Pusateri, W. M. Jay. "Usage of Accessibility Options for the iPhone and iPad in a Visually Impaired Population," *Seminars in ophthalmology*. vol. 32, no. 2, pp. 163–171, 2017.
- [12] Grant, P.; Spencer, L.; Arnoldussen, A.; Hogle, R.; Nau, A.; Szlyk, J.; Nussdorf, J.; Fletcher, D. C.; Gordon, K.; & Seiple, W. "The functional performance of the BrainPort V100 device in persons who are profoundly blind," *Journal of Visual Impairment & Blindness*. vol. 110, no. 2, pp. 77–88.
- [13] J. Spehr, "Object Recognition. In: On Hierarchical Models for Visual Recognition and Learning of Objects, Scenes, and Activities". *Studies in Systems, Decision and Control*, vol 11. Springer, Cham, 2015.
- [14] B. Sandeep, S. Palaniappan, "Kinect language translator by using Google API," *International Journal of Pharmacy and Technology*. vol. 8, no. 4, pp. 20051-20060, 2016.
- [15] J. Campos, S. Colteryahn and K. Gagneja, "IPv6 transmission over BLE Using Raspberry PI 3," 2018. International Conference on Computing, Networking and Communications (ICNC), pp. 200-204.
- [16] M. Richardson and S. Wallace. "Getting Started with Raspberry Pi". O'Reilly Media, Inc., Sebastopol, 2012.
- [17] V. Sati, S. M. Sánchez, N. Shoeibi, A. Arora, and J. M. Corchado, "Face detection and recognition, face emotion recognition through NVIDIA Jetson nano," in *Ambient Intelligence—Software and Applications*, P. Novais, G. Vercelli, J. L. Larriba-Pey, F. Herrera, and P. Chamoso, Eds. Cham, Switzerland: Springer, 2021, pp. 177–185.
- [18] W. A. Indra, A.F.M.F Ismail, Nurulhalim Hassim, M.H. Idris, S.G.Herawan, N.S. Zamzam, F. Zuska., "Feasibility of RF RSL for RF Energy Harvesting : A Case Study of Alor Gajah Area," 2021 IEEE 12th Control and System Graduate Research Colloquium (ICSGRC), 2021, pp. 24-28.
- [19] K. D. Ismael and S. Irina, "Face recognition using Viola-Jones depending on Python," *Indonesian Journal of Electrical Engineering and Computer Science (IJECS)*, vol. 20, no. 3, pp. 1513-1521, December 2020.
- [20] N. Hassim, W. A. Indra, A.F.M.F Ismail, M.S.A.A Majid, S.G.Herawan, N.S. Zamzam, F. Zuska, "GSM900 Downlink Power Density Survey in Merlimau Area," 2021 IEEE 12th Control and System Graduate Research Colloquium (ICSGRC), 2021, pp. 99-103.
- [21] F. Ahmed, D. Tarlow, and D. Batra, "Optimizing expected intersection-over-union with candidate-constrained CRFs," 2015. The IEEE International Conference on Computer Vision, pp.1850-1858.

- [22] W. A. Indra, S. G. Herawan Industrial, N. S. Zamzam, S. b. Mohd Najib Fakulti, N. b. Hassim Fakulti and F. Zuska, "Development of A Guided Drone Powered by Radio Frequency Energy Harvesting," 2021 IEEE International Conference in Power Engineering Application (ICPEA), 2021, pp. 127-131.
- [23] W. A. Indra, A. I. I. Jurjani, N. Hassim, S. G. Herawan, N. S. Zamzam and F. Zuska, "Digital TV Spectrum Survey for the Scope of Energy Scavenging in Jasin, Melaka," 2021 IEEE 17th International Colloquium on Signal Processing & Its Applications (CSPA), 2021, pp. 35-40.
- [24] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (VOC) challenge," International Journal of Computer Vision. vol. 88, no. 2, pp. 303-338, 2010.
- [25] A. -A. Nayan, J. Saha, K. Raqib Mahmud, A. Kalam Al Azad and M. Golam Kibria, "Detection of Objects from Noisy Images," 2020. 2nd International Conference on Sustainable Technologies for Industry 4.0 (STI), pp. 1-6.

Application of Random Forest Regression with Hyper-parameters Tuning to Estimate Reference Evapotranspiration

Satendra Kumar Jain¹

Research Scholar, Department of Computer Science and Applications, Barkatullah University, Bhopal, Madhya Pradesh, India

Anil Kumar Gupta²

Associate Professor, Department of Computer Science and Applications, Barkatullah University, Bhopal, Madhya Pradesh, India

Abstract—Estimation of reference evapotranspiration (ET_o) is a complex and non-linear problem that is used for the quantification of crop water requirements. In this study, random forest regression based models are developed to predict the ET_o of Bhopal city, Madhya Pradesh, India. The meteorological data is collected from IMD, Pune for the periods of the years 2015-16. Based on the correlation among meteorological variables with observed ET_o, four different random forest regression models are created. Moreover, the effects of three important hyper-parameters of random forest, such as the number of trees in the forest, depth of the tree, and the number of samples at a leaf node are evaluated to estimate ET_o using the proposed models. These hyper-parameters are applied in three different ways to the models such as one hyper-parameter parameter at a time, and combination of hyper-parameters using grid search, and random search approaches. In this study, the result indicates that a random forest regression based model with maximal meteorological input variables exhibits great predictive power in small execution time than minimal input variables. This study also reveals that the model that optimises the hyper-parameters using a grid search approach shows equal predictive power but takes much execution time whereas random search based optimization exhibits the same level of predictive capability in less computation time. Stakeholders can utilize random forest regression models with sufficient meteorological data to estimate crop water requirements, and enhance the food production.

Keywords—Reference evapotranspiration; random forest regression; hyper-parameters; grid search; random search optimization

I. INTRODUCTION

Evapotranspiration is a step of the hydrological cycle and has numerous applications such as water management, irrigation scheduling, etc. Evapotranspiration consists of the evaporation and transpiration process. Evaporation removes water from the soils, ponds, and rivers whereas transpiration removes water from the plants. Reference evapotranspiration (ET_o) is estimated on smooth grassland which is further used to estimate crop evapotranspiration. The FAO-PM56 is one of the standard empirical methods provided by the Food Agriculture Organization of the United Nations [1]. Such an empirical method suffers from complicated calculations. Weather stations at various places are equipped with power full devices that are constantly observing climatic data. Machine

learning based models can be applied to such a huge amount of data to estimate ET_o accurately and efficiently. Many authors have applied various machine learning algorithms to estimate ET_o.

The ability of M5P, RF, RT, REPT, and KStar and neuro-fuzzy inference systems such as ANFIS, ANFIS-GA, ANFIS-DE, and ANFIS-ICA has been tested to estimate evapotranspiration [2]. Feed-forward artificial neural network with the Levenberg–Marquardt (LM) training algorithm has been investigated to predict evapotranspiration [3]. Genetic programming (GP), support vector machine-firefly algorithm (SVM-FFA), artificial neural network (ANN), and support vector machine-wavelet (SVM-Wavelet) have been analyzed to predict reference evapotranspiration [4]. Extreme learning machine (ELM), back-propagation neural networks optimized by genetic algorithm (GANN), and wavelet neural networks (WNNgra) models have been developed to estimate evapotranspiration [5]. Random forest (RF) and generalized regression neural network (GRNN) models have been applied to estimate daily evapotranspiration [6]. Four tree based ensemble algorithms such as random forest (RF), M5 model tree (M5Tree), gradient boosting decision tree (GBDT), and extreme gradient boosting (XGBoost) models have been compared for estimation of evapotranspiration [7]. GRNN, MLP, RBNN, GEP, ANFIS-GP, and ANFIS-SC models have been investigated for modeling evapotranspiration [8]. Genetic (GA) and gene expression programming (GEP) models have been used to estimate reference evapotranspiration [9]. M5P Regression Tree, Bagging, Random Forest (RF), and Support Vector Regression (SVR) have been compared [10]. The performance of kNN k-nearest neighbour, artificial neural network, and Adaptive boosting (AdaBoost) to predict daily evaporation for the potato crop have been investigated [11]. Machine learning algorithms have own hyper-parameters that can be tuned at the training duration. Tuning of hyper-parameters can affect the performance of the algorithm. There are various approaches to tune hyper-parameters. In [12] authors show empirically and theoretically that randomly chosen trails are more efficient than the trails on grid. [13] Performed comparative analysis of various hyper-parameters tuning methods to optimize the accuracy of machine learning algorithms. In [14] hyper-parameters are optimized using weighted random search approaches.

Estimation of ETo plays an important role in water saving and enhancement of food production leading to food security in the world. The selection of machine learning algorithms to estimate ETo is a challenging task because they are not good for all problems. The size and structures of the data affect the performance of the machine learning algorithm. In the current study, the random forest regression algorithm is chosen because of its high performance and handling a complex problem. In this paper, the contribution of works is summarized as follows:

- The reviews of machine learning techniques to estimate reference evapotranspiration and hyper-parameter tuning approaches are done.
- Meteorological data of Bhopal city is collected from IMD Pune. Descriptive analysis is performed on preprocessed data. The correlation coefficient of meteorological data with observed ETo is determined.
- Four different random forest regression based data driven models (based on the correlation among meteorological variables with observed ETo) are developed.
- These hyper-parameters (n_estimators, max_depth, min_samples_leaf) are applied in three different ways ('one parameter at a time, combinations of parameters using grid and random search approaches) to the four random forest models.
- The performances of twenty models are evaluated and compared with FAO-PM56 using six statistical indicators.

II. MATERIAL AND METHODS

A. Study Site

The proposed random forest regression based data driven models are analyzed in this study using meteorological data from Bhopal city of Madhya Pradesh state, India. Daily meteorological data for the years 2015-16 are obtained from the India Meteorological Data, Pune, which includes input attributes such as minimum temperature (Tmin) in 0C, maximum temperature (Tmax) in 0C, relative humidity (RH) in %, wind speed (u) in m/s and mean solar radiation (Rn) in MJ m-2 day-1. Daily mean sunshine hours of Bhopal city are taken from the Daily Normals of Global & Diffuse Radiation report issued by IMD Pune published in the year 2016. Bhopal city has a subtropical humid climate. It has an average elevation of 500 meters and is located at 23.25 oE latitude and 77.42 oN longitude. Descriptions of training and test datasets are summarized in Table I. The monthly variation of ETo at Bhopal city is observed, where the average minimum ETo is 2.33 mm/day in January 2015 and 2.27 mm/day in December 2016 is noted, similarly average maximum ETo is 7.0 mm/day in May 2015 and 7.47 mm/day in May 2016 is noted. The correlation matrix of observed ETo and the meteorological data of Bhopal city is given in Table II. It can be observed that ETo has a positive correlation with temperature, solar radiation, and wind speed parameters whereas a negative correlation with humidity. Hence it can be said that ETo is an energy driven

process and increases as temperature, radiation, and wind speed are increased.

B. FAO-PM56 Equation

The FAO-56 Penman-Monteith equation is provided by the Food and Agriculture Organization of the United Nation and is considered a standard worldwide accepted method to estimate ETo. It is represented as-

$$ETo = \frac{0.408 * (R_n - G) + \gamma \left(\frac{900}{T + 273} \right) u_2 (e_s - e_a)}{\Delta + \gamma (1 + 0.34 * u_2)} \quad (1)$$

Where

ETo = grass reference evapotranspiration in mm day⁻¹

R_n = net radiation in MJ mm⁻² day⁻¹

G = soil heat flux in MJ mm⁻² day⁻¹

γ = psychrometric constant in kPa⁰ c⁻¹

T = mean daily air temperature in °C

u_2 = wind speed at 2m height in m s⁻¹

Δ = slope vapour pressure curve in kPa⁰ c⁻¹

e_s = saturation vapour pressure in kPa

e_a = actual vapour pressure in kPa

$e_s - e_a$ = saturation vapour pressure deficit in kPa

TABLE I. STATISTICAL DESCRIPTION OF METEOROLOGICAL DATA

Climatic parameters	Data set	Minimum	Maximum	Mean	Standard Deviation
T _{min}	Training	5.8	32.1	19.63	6.0
	Test	7.9	31.2	20.25	5.75
T _{max}	Training	15.5	46.7	32.17	5.8
	Test	14.2	45.3	32.84	5.74
RH	Training	12	99	56.85	21.76
	Test	17	98	55.19	23.04
u	Training	0	6.6	1.01	0.65
	Test	0.2	2.9	0.99	0.57
R _n	Training	12.3	26.3	18.83	3.48
	Test	12.3	26.3	19.11	3.65
ETo	Training	1.71	9.5	4.06	1.57
	Test	1.56	8.1	4.15	1.58

TABLE II. CORRELATION COEFFICIENT OF METEOROLOGICAL DATA WITH OBSERVED ETO

	T _{min}	T _{max}	RH	u	R _n	ETo
T _{min}	1					
T _{max}	0.72	1				
RH	-0.054	-0.6	1			
u	0.49	0.22	0.09	1		
R _n	0.36	0.70	-0.74	0.14	1	
ETo	0.71	0.87	-0.60	0.47	0.82	1

ETo is observed by CROPWAT8.0 software in this study, which is a decision support tool and developed by the Land and Water Development division of the Food and Agriculture Organization of the United Nation. Daily minimum temperature (T_{\min}), maximum temperature (T_{\max}), relative humidity (R_H), bright sunshine hours (I_s), and wind speed (u) are applied as input parameters to CROPWAT8.0 software and it returns daily or monthly solar radiation (R_n) and ETo (mm day^{-1}). The FAO-PM56 is considered superior to other methods if reliable and complete meteorological data are available. Huge amounts of meteorological data are recorded at weather stations. Estimation of ETo from such large data using machine learning based models could be an alternative solution that produces accurate and efficient outcomes.

C. Random Forest Regression

Random forest is a supervised machine learning algorithm that is used for classification as well as regression problems. In this study a random forest machine learning algorithm is used to estimate ETo of Bhopal city, which is considered as a function approximation (regression) problem. It works based on the ensemble learning concept, in which instead of making a single model, multiple models are created on randomly selected data. Therefore the outcome of the random forest regression is made based on estimated results of multiple models [15]. Hence it is considered a highly stable model. It removes the overfitting problem of a decision tree. Multiple trees in the random forest lead to higher accuracy. It works well for large datasets with high dimensions. Various hyper-parameters are provided for the random forest. Tuning of hyper-parameters may improve the performance and predictive capability of random forests. Number of trees in the forest ($n_{\text{estimators}}$), the longest path between the root and the leaf node (max_depth), the minimum required samples to split a node in the tree (min_samples_split), the maximum number of leaf nodes in the tree (max_leaf_nodes), minimum number of samples at the leaf nodes (min_samples_leaf), and criteria to split the node in the tree (criterion) are considered some important hyper-parameters of random forest. In the present study, the performance of random forest is evaluated by tuning the three hyper-parameters such as $n_{\text{estimators}}$ (10, 20, 30, ..., 100), max_depth (2, 3, 4, ..., 10), and min_samples_leaf (2, 3, 4, 5). These hyper-parameters are applied in three different ways to the models such as 'one hyper-parameter at a time', and 'combinations of hyper-parameters' using grid search, and random search approaches. In the case of 'one hyper-parameter at a time', the search space consists of one dimensional hyper parameter values. Grid search and random search approaches are used when multiple hyper-parameters are applied to the model. In this case, the search space consists of a grid of hyper-parameter values, and the model is evaluated at each point in the grid. In the case of random search, the model is evaluated on a randomly opted grid point. Grid search is simple to implement and always finds the best combinations of hyper-parameter. It is a time consuming approach due to the exhaustive search nature. Random search exhibits the same performance in less computation time.

D. Model Development

Model development steps are shown in Fig. 1. Initially, the meteorological and geographical data of Bhopal city is taken

into memory. Data preprocessing is a significant step to estimate ETo accurately. It transforms the data in a meaningful way. To obtain the optimum outcomes, missing values are filled in different ways. In the present study, missing values are filled by the mean value of those attributes. Values of all attributes are normalized using the z-score method to make all attributes to the same level of magnitudes so have the same emphasis. Values of ETo are observed using CROPWAT 8.0 software (developed by the Land and Water Development Division of FAO (The Food and Agriculture Organization of the United Nation)) and made as a dependent variable, whereas the remaining attributes (T_{\min} , T_{\max} , R_n , u , R_H) are designated as independent variables. The whole dataset is partitioned into the training dataset (80%) and the test dataset (20%).

Four random forest regression based models such as RFR-Model1, RFR-Model2, RFR-Model3, and RFR-Model4 are created. Different combinations of meteorological input parameters (made based on high correlation coefficient with observed ETo values) are applied to these models. In the RFR-Model1, T_{\min} , and T_{\max} are applied. In the RFR-Model2, T_{\min} , T_{\max} , and R_n are applied. In the RFR-Model3, T_{\min} , T_{\max} , R_n , and u are applied. And finally in the RFR-Model4, T_{\min} , T_{\max} , R_n , u , and R_H are applied. In addition to the input combinations of meteorological parameters, three important hyper-parameters are tuned in each model. These hyper-parameters are tuned and applied to the proposed four models in three different ways: 'one hyper-parameter at a time', and combinations of hyper-parameters are using grid search, and random search optimization approach. Taking into consideration four different models and the applicability of three hyper-parameters to the models produces twenty combinations. Therefore in this study, the performances of twenty models are evaluated. Six different statistical indicators are used in this study to evaluate the performance of the models such as mean absolute error (MAE), mean squared error (MSE), root mean squared error (RMSE), Pearson correlation coefficient (r), r^2 (coefficient of determination), and Nash-Sutcliffe(NS). These models are implemented in Python with the help of Pandas, Numpy, Sklearn and Matplotlib libraries.

E. Performance Evaluation Indices

Predictive skills of RFR-MODEL1, RFR-MODEL2, RFR-MODEL3, and RFR-MODEL4 are evaluated using the following parameters:

Mean absolute error (MAE).

$$MAE = \frac{\sum_{i=1}^n |E_{pi} - E_{oi}|}{n} \quad (2)$$

where

E_{pi} = predicted evapotranspiration.

E_{oi} = observed evapotranspiration.

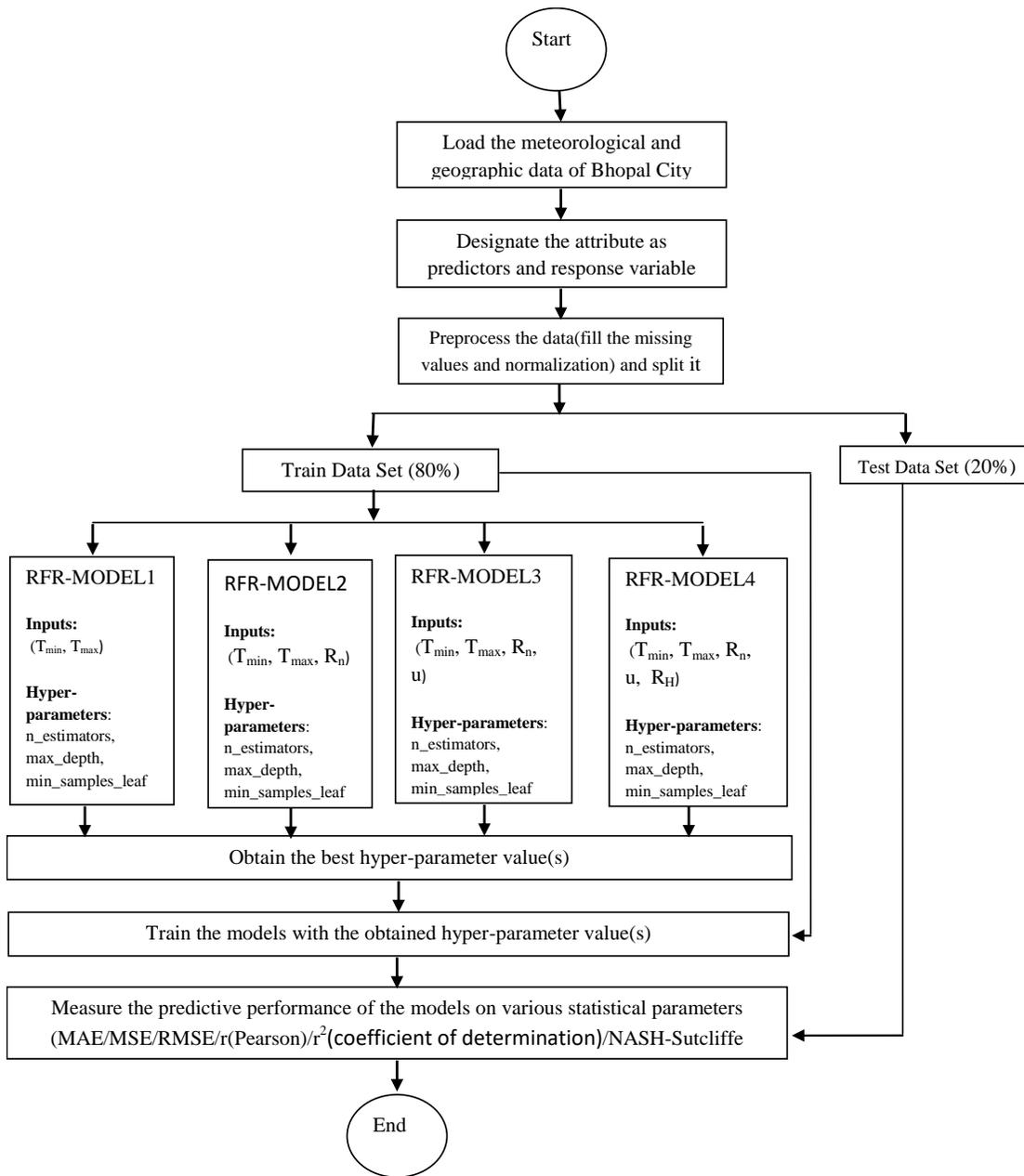


Fig. 1. Flow Chart of Random Forest Regression based Models.

Mean square error (MSE).

$$MSE = \frac{1}{n} \sum_{i=1}^n (E_{oi} - E_{pi})^2 \quad (3)$$

Root mean square error (RMSE) -A small value of RMSE denotes the model fits the datasets strongly.

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (E_{oi} - E_{pi})^2} \quad (4)$$

Pearson correlation coefficient shows the strength of the relationship between observed and predicted ET_o.

$$Pearson \text{ correlation}(r) = \frac{\sum_{i=1}^n (E_{pi} - \overline{E_{pi}})(E_{oi} - \overline{E_{oi}})}{\sqrt{\sum_{i=1}^n (E_{pi} - \overline{E_{pi}})^2 \sum_{i=1}^n (E_{oi} - \overline{E_{oi}})^2}} \quad (5)$$

r^2 (coefficient of determination) . A larger value of r^2 indicates the model fits the datasets strongly.

$$Coefficient \text{ of det er min ation}(r^2) = r * r \quad (6)$$

Nash-Sutcliffe efficiency (NS) is used to assess the predictive skills of ANN models.

$$Nash - Sutcliffe\ efficiency(NS) = 1 - \frac{\sum_{i=1}^n (E_{pi} - E_{oi})^2}{\sum_{i=1}^n (E_{pi} - \bar{E}_o)^2} \quad (7)$$

III. RESULT AND DISCUSSION

As stated earlier in the model development section, taking into consideration four different random forest regression based models and the applicability of three hyper-parameters to the models in different ways produces twenty combinations. Therefore in this study, the performances of twenty models are evaluated. The execution time span of each model is calculated from the beginning of the training period to the end of the testing period.

A. Performance of the RFR-Model1

In this model, only two meteorological inputs T_{min} , and T_{max} are applied. The performance of this model is demonstrated in Table III, where it exhibits mae of 0.48, mse of 0.39, rmse of 0.62, r of 0.92, r^2 of 0.85, and Nash-Sutcliffe of 0.85 when the $n_estimators$ hyper-parameter is tuned. Table IV exhibits the performance with mae of 0.45, mse of 0.34, rmse of 0.59, r of 0.93, r^2 of 0.87, and Nash-Sutcliffe of 0.86 when the max_depth hyper-parameter is tuned. Table V exhibits the performance with mae of 0.45, mse of 0.35, rmse of 0.59, r of 0.93, r^2 of 0.87, and Nash-Sutcliffe of 0.86 when the $min_samples_leaf$ hyper-parameter is tuned. Table VI exhibits the performance with mae of 0.46, mse of 0.36, rmse of 0.6, r of 0.93, r^2 of 0.86, and Nash-Sutcliffe of 0.86 when the combination of three hyper-parameters ($n_estimators$, max_depth , $min_samples_leaf$) are tuned using a grid search approach. Similarly, Table VII exhibits the performance with mae of 0.46, mse of 0.35, rmse of 0.59, r of 0.93, r^2 of 0.87, and Nash-Sutcliffe of 0.86 when the same combination of hyper-parameters are tuned using a random search approach. It can be observed that RFR-Model1 shows almost the same predictive capability in all scenarios. The Computation time of this model is represented in Table VIII. It takes 10.5 seconds when the $n_estimators$ hyper-parameter is tuned, 29.38 seconds when the max_depth hyper-parameter is tuned, 70 seconds when the $min_samples_leaf$ hyper-parameter is tuned, 301.33 seconds when a grid search approach is applied, and 11.62 seconds when a random search approach is applied respectively in order to estimate ETo. Regression analysis of the RFR-Model1 is shown in Fig. 2 for all scenarios.

B. Performance of the RFR-Model2

In this model, only three meteorological inputs T_{min} , T_{max} , and R_n are applied. The performance of this model is demonstrated in Table III, where it exhibits mae of 0.29, mse of 0.17, rmse of 0.41, r of 0.97, r^2 of 0.93, and Nash-Sutcliffe of 0.93 when the $n_estimators$ hyper-parameter is tuned. Table IV exhibits the performance with mae of 0.29, mse of 0.17, rmse of 0.41, r of 0.97, r^2 of 0.94, and Nash-Sutcliffe of 0.93 when the max_depth hyper-parameter is tuned. Table V exhibits the performance with mae of 0.29, mse of 0.17, rmse of 0.41, r of 0.97, r^2 of 0.94, and Nash-Sutcliffe of 0.93 when the $min_samples_leaf$ hyper-parameter is tuned. Table VI exhibits the performance with mae of 0.29, mse of 0.17, rmse

of 0.41, r of 0.97, r^2 of 0.94, and Nash-Sutcliffe of 0.93 when the combination of three hyper-parameters ($n_estimators$, max_depth , $min_samples_leaf$) are tuned using a grid search approach. Similarly, Table VII exhibits the performance with mae of 0.29, mse of 0.17, rmse of 0.41, r of 0.97, r^2 of 0.94, and Nash-Sutcliffe of 0.93 when the same combination of hyper-parameters are tuned using a random search approach. It can be observed that RFR-Model2 shows the same predictive capability in all scenarios but higher than RFR-Model1. The Computation time of this model is represented in Table VIII. It takes 11.62 seconds when the $n_estimators$ hyper-parameter is tuned, 31.9 seconds when the max_depth hyper-parameter is tuned, 69 seconds when the $min_samples_leaf$ hyper-parameter is tuned, 321.39 seconds when a grid search approach is applied, and 9.72 seconds when a random search approach is applied respectively in order to estimate ETo. Regression analysis of the RFR-Model2 is shown in Fig. 3 for all scenarios.

TABLE III. MODEL PERFORMANCE WHEN 'N_ESTIMATORS' HYPER PARAMETER IS TUNED

Performance Indices	RFR-Model 1	RFR-Model 2	RFR-Model 3	RFR-Model 4
MAE	0.48	0.29	0.17	0.15
MSE	0.39	0.17	0.05	0.05
RMSE	0.62	0.41	0.23	0.22
Pearson(r)	0.92	0.97	0.99	0.99
r^2	0.85	0.93	0.98	0.98
Nash-Sutcliffe	0.85	0.93	0.98	0.98

TABLE IV. MODEL PERFORMANCE WHEN 'MAX_DEPTH' HYPER PARAMETER IS TUNED

Performance Indices	RFR-Model1	RFR-Model2	RFR-Model3	RFR-Model4
MAE	0.45	0.29	0.17	0.15
MSE	0.34	0.17	0.05	0.05
RMSE	0.59	0.41	0.23	0.22
Pearson(r)	0.93	0.97	0.99	0.99
r^2	0.87	0.94	0.98	0.98
Nash-Sutcliffe	0.86	0.93	0.98	0.98

TABLE V. MODEL PERFORMANCE WHEN 'MAX_SAMPLES_LEAF' HYPER PARAMETER IS TUNED

Performance Indices	RFR-Model1	RFR-Model2	RFR-Model3	RFR-Model4
MAE	0.45	0.29	0.18	0.16
MSE	0.35	0.17	0.06	0.06
RMSE	0.59	0.41	0.25	0.23
Pearson(r)	0.93	0.97	0.99	0.99
r^2	0.87	0.94	0.98	0.98
Nash-Sutcliffe	0.86	0.93	0.97	0.98

TABLE VI. MODEL PERFORMANCE WHEN (N_ESTIMATORS, MAX_DEPTH, MAX_SAMPLES_LEAF) HYPER PARAMETERS ARE TUNED USING GRID SEARCH

Performance Indices	RFR-Model1	RFR-Model2	RFR-Model3	RFR-Model4
MAE	0.46	0.29	0.18	0.17
MSE	0.36	0.17	0.06	0.06
RMSE	0.6	0.41	0.25	0.24
Pearson(r)	0.93	0.97	0.99	0.99
r ²	0.86	0.94	0.98	0.98
Nash-Sutcliffe	0.86	0.93	0.97	0.98

TABLE VII. MODEL PERFORMANCE WHEN N_ESTIMATORS, MAX_DEPTH, MAX_SAMPLES_LEAF) HYPER PARAMETERS ARE TUNED USING RANDOM SEARCH

Performance Indices	RFR-Model1	RFR-Model2	RFR-Model3	RFR-Model4
MAE	0.46	0.29	0.19	0.16
MSE	0.35	0.17	0.07	0.06
RMSE	0.59	0.41	0.26	0.24
Pearson(r)	0.93	0.97	0.99	0.99
r ²	0.87	0.94	0.98	0.98
Nash-Sutcliffe	0.86	0.93	0.97	0.98

TABLE VIII. EXECUTION TIME (SECONDS) TAKEN BY EACH MODEL

Models	One hyper-parameter at time			grid search	random search
	number of trees	depth of trees	samples at leaf node		
RFR-Model1	10.5	29.38	70	301.33	11.62
RFR-Model2	11.62	31.9	69	321.39	9.72
RFR-Model3	11.87	33.24	71.2	316.31	9.48
RFR-Model4	12.44	36.8	75.57	329.98	11.12

C. Performance of the RFR-Model3

In this model, four meteorological inputs T_{min} , T_{max} , R_n , and u are applied. The performance of this model is demonstrated in Table III, where it exhibits mae of 0.17, mse of 0.05, rmse of 0.23, r of 0.99, r² of 0.98, and Nash-Sutcliffe of 0.98 when the n_estimators hyper-parameter is tuned. Table IV exhibits the performance with mae of 0.17, mse of 0.05, rmse of 0.23, r of 0.99, r² of 0.98, and Nash-Sutcliffe of 0.98 when the max_depth hyper-parameter is tuned. Table V exhibits the performance with mae of 0.18, mse of 0.06, rmse of 0.25, r of 0.99, r² of 0.98, and Nash-Sutcliffe of 0.97 when the min_samples_leaf hyper-parameter is tuned. Table VI exhibits the performance with mae of 0.18, mse of 0.06, rmse of 0.25, r of 0.99, r² of 0.98, and Nash-Sutcliffe of 0.97 when the combination of three hyper-parameters (n_estimators, max_depth, min_samples_leaf) are tuned using a grid search approach. Similarly, Table VII exhibits the performance with mae of 0.19, mse of 0.07, rmse of 0.26, r of 0.99, r² of 0.98, and

Nash-Sutcliffe of 0.97 when the same combination of hyper-parameters are tuned using a random search approach. It can be observed that RFR-Model3 shows almost the same predictive capability with minor variations in all scenarios but higher than RFR-Model1 and RFR-Model2. The Computation time of this model is represented in Table VIII. It takes 11.87 seconds when the n_estimators hyper-parameter is tuned, 33.24 seconds when the max_depth hyper-parameter is tuned, 71.2 seconds when the min_samples_leaf hyper-parameter is tuned, 316.31 seconds when a grid search approach is applied, and 9.48 seconds when a random search approach is applied respectively in order to estimate ETo. Regression analysis of the RFR-Model3 is shown in Fig. 4 for all scenarios.

D. Performance of the RFR-Model4

In this model, five meteorological inputs T_{min} , T_{max} , R_n , u , and R_H are applied. The performance of this model is demonstrated in Table III, where it exhibits mae of 0.15, mse of 0.05, rmse of 0.22, r of 0.99, r² of 0.98, and Nash-Sutcliffe of 0.98 when the n_estimators hyper-parameter is tuned. Table IV exhibits the performance with mae of 0.15, mse of 0.05, rmse of 0.22, r of 0.99, r² of 0.98, and Nash-Sutcliffe of 0.98 when the max_depth hyper-parameter is tuned. Table V exhibits the performance with mae of 0.16, mse of 0.06, rmse of 0.23, r of 0.99, r² of 0.98, and Nash-Sutcliffe of 0.98 when the min_samples_leaf hyper-parameter is tuned. Table VI exhibits the performance with mae of 0.17, mse of 0.06, rmse of 0.24, r of 0.99, r² of 0.98, and Nash-Sutcliffe of 0.98 when the combination of three hyper-parameters (n_estimators, max_depth, min_samples_leaf) are tuned using a grid search approach. Similarly, Table VII exhibits the performance with mae of 0.16, mse of 0.06, rmse of 0.24, r of 0.99, r² of 0.98, and Nash-Sutcliffe of 0.98 when the same combination of hyper-parameters are tuned using a random search approach. It can be observed that RFR-Model4 shows almost the same predictive capability in all scenarios but higher than RFR-Model1, RFR-Model2 and RFR-Model3. The Computation time of this model is represented in Table VIII. It takes 12.44 seconds when the n_estimators hyper-parameter is tuned, 36.8 seconds when the max_depth hyper-parameter is tuned, 75.57 seconds when the min_samples_leaf hyper-parameter is tuned, 329.98 seconds when a grid search approach is applied, and 11.12 seconds when a random search approach is applied respectively in order to estimate ETo. Regression analysis of the RFR-Model4 is shown in Fig. 5 for all scenarios.

It can be observed that RFR-Model1 demonstrates poor predictive performance. The performance of the models is improving gradually when the maximal meteorological input variables are taken into consideration. Grid search based optimization demonstrates the same level of performance but takes much execution time and will not be feasible when size of search spaces increases whereas random search based optimization exhibits better performance than grid search. Computation time is shown in Fig. 6.

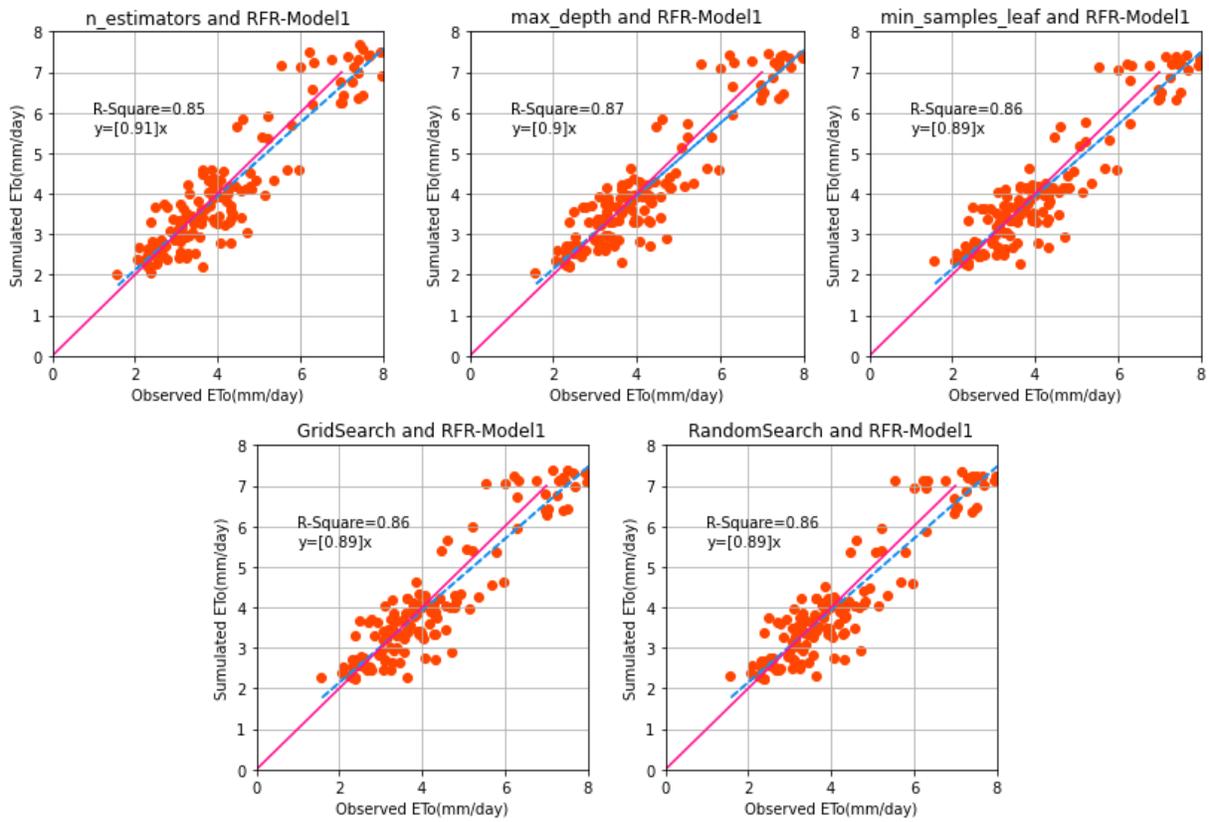


Fig. 2. Regression Analysis of the RFR-Model1 in all Scenarios.

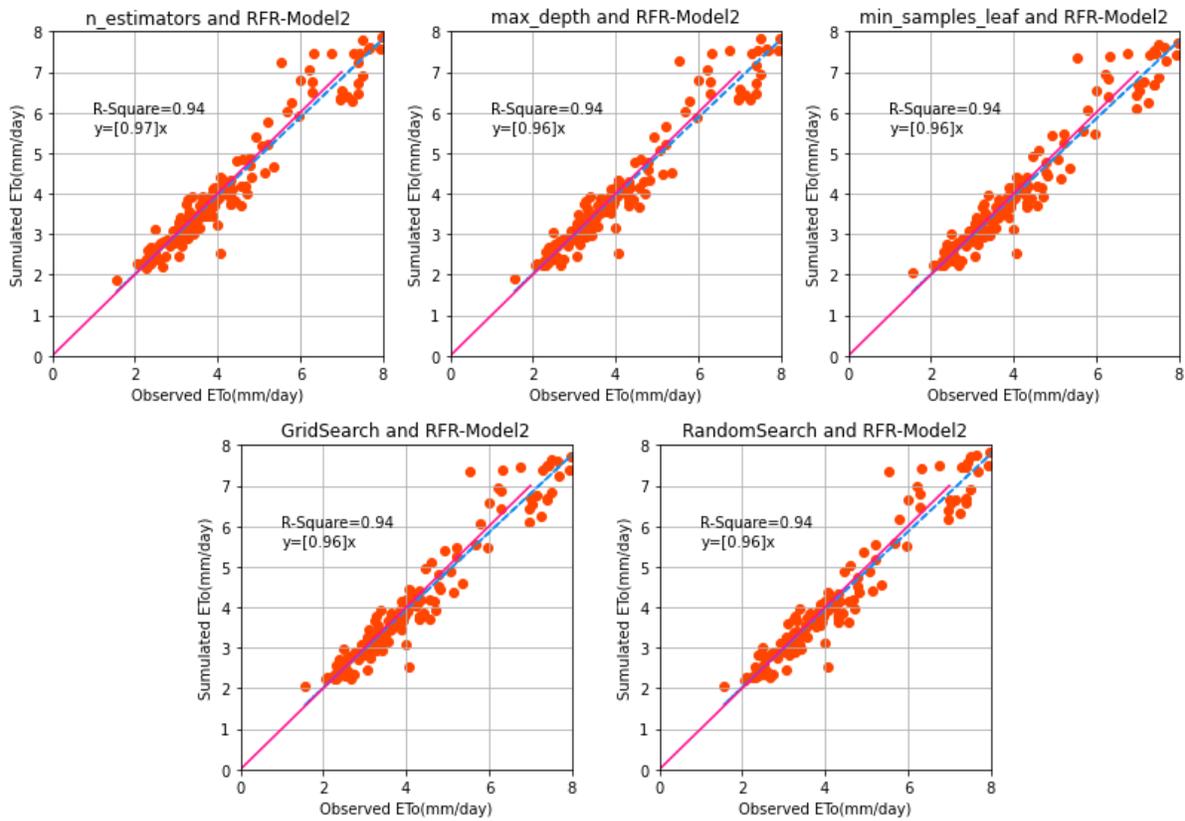


Fig. 3. Regression Analysis of the RFR-Model2 in all Scenarios.

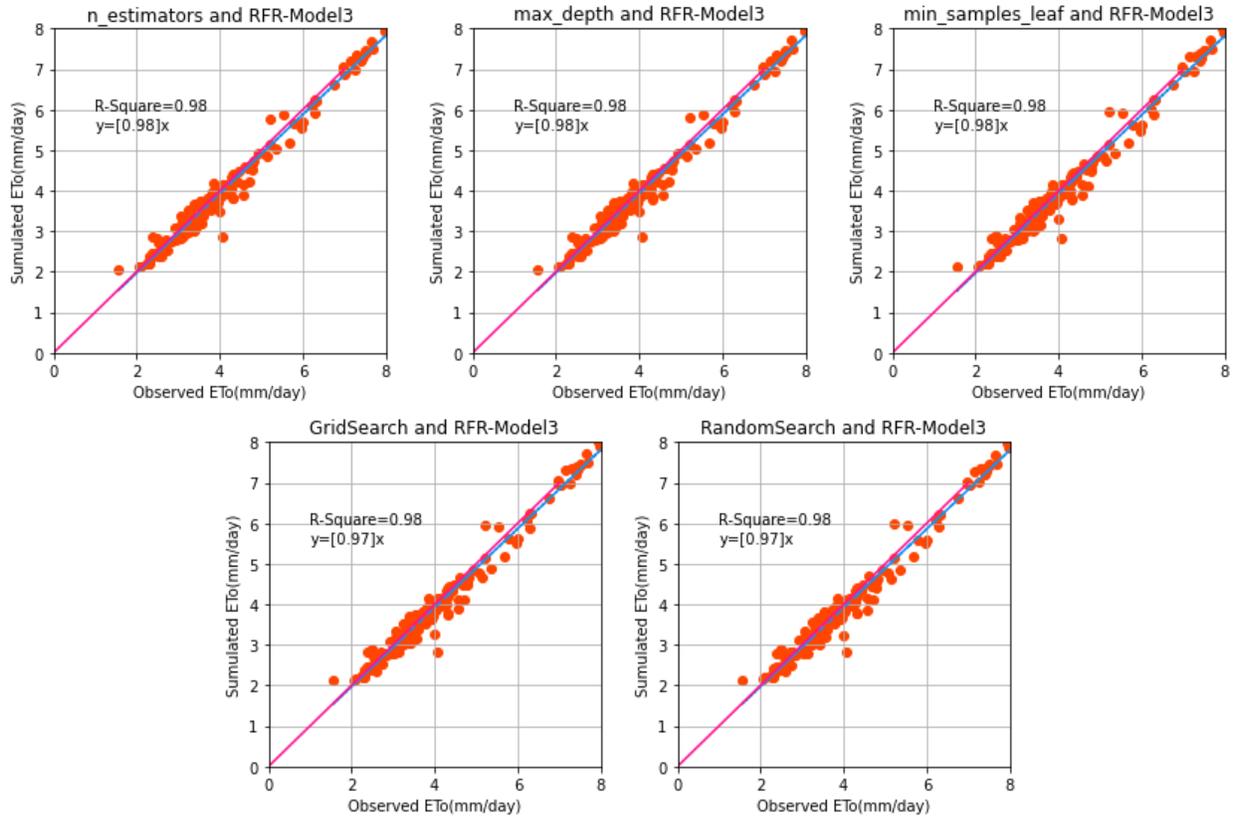


Fig. 4. Regression Analysis of the RFR-Model3 in all Scenarios.

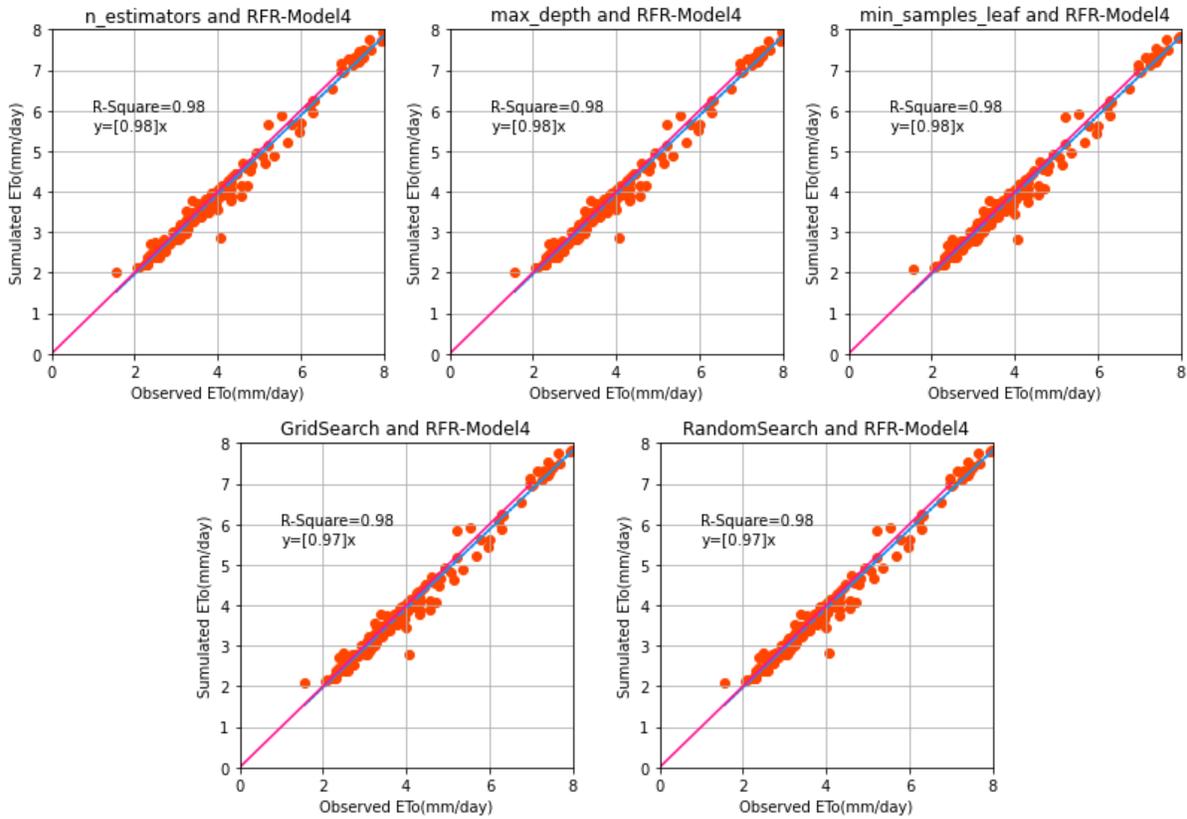


Fig. 5. Regression Analysis of the RFR-Model4 in all Scenarios.

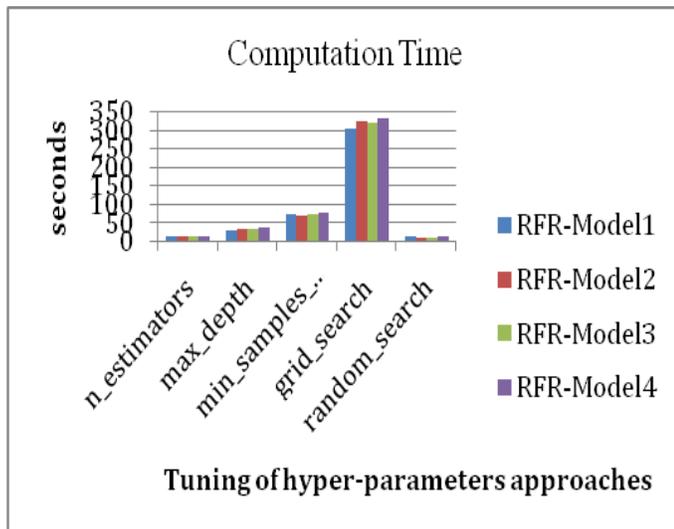


Fig. 6. Computation Time of each Model.

IV. CONCLUSION

Estimation of ETo has numerous applications. Irrigation scheduling is one of them. In this study, random forest regression based four different models are developed to estimate ETo. Different combinations of meteorological input variables (made based on high correlation coefficient with observed ETo values) are applied to these models. Moreover, the effects of three important hyper-parameters of random forest regression, such as the number of trees in the forest, depth of the trees, and the number of samples at a leaf node are evaluated to estimate ETo using the proposed models. These hyper-parameters are optimized and applied in three different ways to the models such as one parameter at a time, and combinations of hyper parameters using grid search, and random search. This study reveals that the models with less meteorological input variables demonstrate poor performance than models with maximal input variables (r is of 0.99, r^2 is of 0.98 and Nash-Sutcliffe is of 0.98 in the case of RFR-Model4). Models based on grid search based optimization exhibit the same predictive power but take much computation time. The findings of this study are that random forest regression based models with sufficient meteorological data demonstrate better performance and are useful to the stakeholders such as farmers, engineers for irrigation scheduling and water management. In the future, more hyper-parameter optimization techniques will be applied to estimate accurate ETo for various places in India. This estimated ETo will be used to calculate crop water requirements of Wheat and Maize crops

REFERENCES

- [1] R. G. Allen, L. S. Pereira, D. Raes, and M. Smith, "Crop evapotranspiration-Guidelines for computing crop water requirements-FAO Irrigation and drainage paper 56," 1998.
- [2] K. Khosravi et al., "Meteorological data mining and hybrid data-intelligence models for reference evaporation simulation: A case study in Iraq," *Comput. Electron. Agric.*, vol. 167, Dec. 2019, doi: 10.1016/j.compag.2019.105041.
- [3] O. Kisi, "Evapotranspiration modelling from climatic data using a neural computing technique," *Hydrol. Process.*, vol. 21, no. 14, pp. 1925–1934, Jul. 2007, doi: 10.1002/hyp.6403.
- [4] M. Gocić et al., "Soft computing approaches for forecasting reference evapotranspiration," *Comput. Electron. Agric.*, vol. 113, pp. 164–173, Apr. 2015, doi: 10.1016/j.compag.2015.02.010.
- [5] Y. Feng, N. Cui, L. Zhao, X. Hu, and D. Gong, "Comparison of ELM, GANN, WNN and empirical models for estimating reference evapotranspiration in humid region of Southwest China," *J. Hydrol.*, vol. 536, pp. 376–383, May 2016, doi: 10.1016/j.jhydrol.2016.02.053.
- [6] Y. Feng, N. Cui, D. Gong, Q. Zhang, and L. Zhao, "Evaluation of random forests and generalized regression neural networks for daily reference evapotranspiration modelling," *Agric. Water Manag.*, vol. 193, pp. 163–173, Nov. 2017, doi: 10.1016/j.agwat.2017.08.003.
- [7] J. Fan et al., "Evaluation of SVM, ELM and four tree-based ensemble models for predicting daily reference evapotranspiration using limited meteorological data in different climates of China," *Agric. For. Meteorol.*, vol. 263, pp. 225–241, Dec. 2018, doi: 10.1016/j.agrformet.2018.08.019.
- [8] H. Sanikhani, O. Kisi, E. Maroufpoor, and Z. M. Yaseen, "Temperature-based modeling of reference evapotranspiration using several artificial intelligence models: application of different modeling scenarios," *Theor. Appl. Climatol.*, vol. 135, no. 1–2, pp. 449–462, Jan. 2019, doi: 10.1007/s00704-018-2390-z.
- [9] M. Valipour, M. A. G. Sefidkouhi, M. Raeini-Sarjaz, and S. M. Guzman, "A hybrid data-driven machine learning technique for evapotranspiration modeling in various climates," *Atmosphere (Basel)*, vol. 10, no. 6, Jun. 2019, doi: 10.3390/atmos10060311.
- [10] F. Granata, "Evapotranspiration evaluation models based on machine learning algorithms—A comparative study," *Agric. Water Manag.*, vol. 217, pp. 303–315, May 2019, doi: 10.1016/j.agwat.2019.03.015.
- [11] S. S. Yamaç and M. Todorovic, "Estimation of daily potato crop evapotranspiration using three different machine learning algorithms and four scenarios of available meteorological data," *Agric. Water Manag.*, vol. 228, Feb. 2020, doi: 10.1016/j.agwat.2019.105875.
- [12] J. Bergstra, J. B. Ca, and Y. B. Ca, "Random Search for Hyper-Parameter Optimization Yoshua Bengio," 2012. [Online]. Available: <http://scikit-learn.sourceforge.net>.
- [13] E. Elgeldawi, A. Sayed, A. R. Galal, and A. M. Zaki, "Hyperparameter tuning for machine learning algorithms used for arabic sentiment analysis," *Informatics*, vol. 8, no. 4, Dec. 2021, doi: 10.3390/informatics8040079.
- [14] R. Andonie and A. C. Florea, "Weighted random search for CNN hyperparameter optimization," *Int. J. Comput. Commun. Control*, vol. 15, no. 2, pp. 1–11, 2020, doi: 10.15837/IJCCC.2020.2.3868.
- [15] L. Breiman, "Random Forests," 2001.

Abnormal Event Detection using Additive Summarization Model for Intelligent Transportation Systems

G. Balamurugan¹

Research Scholar

Department of Computer Science and Engineering
Puducherry Technological University, Pondicherry, India

Dr. J. Jayabharathy²

Associate Professor

Department of Computer Science and Engineering
Puducherry Technological University, Pondicherry, India

Abstract—Video surveillance is used for capturing the abnormal events on roadsides that are caused due to improper driving, accidents, and hindrances resulting in transportation lags and life-critical issues. It is essential to highlight the accident keyframes in videos to achieve intelligent video surveillance. Video summarization plays a vital role in summarizing the keyframe for an abnormal event from the stacked video surveillance input. The observed video is converted into frames and analyzed for providing an accurate summarization for accident analysis forecast and guiding the users in avoiding such events. The main issues in summarization arise from the inconsistency between the spatiotemporal redundancies and the classification of sequence verification in video surveillance. This article introduces an Additive Event Summarization Method (AESM) for projecting classified events through a gated recurrent unit learning paradigm. In this process, the gates are assigned for unclassified and active frames for sequence verification. Based on the sequence, the abnormality is classified and summarized with higher accuracy than the state of art techniques. This proposed method relies on heterogeneous features for classifying events with better structural indices. The proposed method's performance is analyzed using the metrics accuracy, false rate, analysis time, SSIM, and F1-Score.

Keywords—Event detection; gated recurrent unit; summarization; intelligent transportation system

I. INTRODUCTION

Road transportation is one of the cheapest and easiest among the other types of the transportation system. Many people around the world are traveling via road to travel from one place to another. Abnormal event detection is one of the critical tasks to perform in transportation systems [1]. Closer circuit television (CCTV) plays a major role in detecting events which are occurred on the roadside. A transport monitoring system plays a vital role in analyzing every detail which is occurred on the roadside and helps to protect people. Accurate abnormal event detection helps to reduce the crime rate and death rate on roadsides [2]. Abnormal events on roadsides are classified based on the physical attributes and behavior, postures, and gestures of the vehicle's position. The abnormal Event detection process is done based on two stages namely classification and video summarization process [3].

The video summarization process is done by analyzing the events which are occurred on the roadside based on certain

keyframes or parameters from the given video clips. Keyframe plays a major role in the summarization process which helps to identify the exact features of the video which is done by comparing it with an important set of features [4]. The classification process is processed by combining both normal and abnormal events which are occurred on the roadside and then it produced a dataset that contains the cause of abnormal events in a detailed manner [5]. The Video Summarization process is used in every monitoring system to enhance the network by understanding the exact cause of events by analyzing the given set of videos [6]. The video summarization process helps to control accidents and crime on roadsides. A keyframe is generated to identify the exact actual cause of the events and it also helps to find out the upcoming events based on people's activities [7]. The machine learning algorithm is mostly used in the summarization process which helps to increase the accuracy rate in the detection process and also helps to reduce the time consumption rate in processing data [8]. A dynamic hierarchical clustering algorithm is used in the summarization process which is done by training the data which are captured by CCTV and producing trained data for further uses. It is done by combining the current clips or data with the previously collected data and generating detailed information which helps to prevent the upcoming accident [3, 9]. A reinforcement algorithm is also used here to identify the keyframes based on features such as gestures, signs, and postures of people and produce sequenced keyframes which help to reduce the crime rate on roadsides [10]. The main disadvantages of the video summarization cause contradiction between the spatiotemporal redundancies and sequence prediction. The proposed Additive Event Summarization Method (AESM) is used for projecting classified events through a gated recurrent unit. The proposed system is used to increase the chances of early classification due to limited state-based classification and summarization. The main advantages of the proposed system are to decrease the computational complexity caused by recurrent replication-based classification and reduce the sensitivity of the output. The experimental analysis of the proposed work is conducted using the dataset of UCSD to find the predominance compared to state of art techniques. The remaining sections of the paper are organized as follows. Section 2 presents the analysis of the related work with the merits and limitations. Section 3 presents the proposed work with a detailed mathematical

analysis of the summarization and classification process. Section 4 describes the experimental analysis of the proposed work. Section 5 concludes the paper with its contributions and the scope of the research.

II. RELATED WORK

Yang et al. [11] proposed an algorithm for the learning model in the real-time event summarization process HRES. The learning model is proposed to capture the information which is stored in the knowledge base (KB) and implicitly the information based on the queries which are given to the users. The proposed HRES method improves the robustness and effectiveness and reduces the time consumption rate.

Wan et al. [12] proposed a long video retrieval algorithm based on a superframe segmentation process for ITS event detection. A long video stream is used to identify the unwanted frames which are present in the database and helps to reduce the unnecessary frame. The segment of Interest (SOI) is generated by using the superframe segmentation process. The proposed method increases the effectiveness by reducing the retrieval time.

Thomas et al. [13] proposed video summarization based on a perceptual model for the roadside event detection process. This method is used to find out the optimal solutions by analyzing the vast number of videos that are captured during accidents time. The surveillance camera is used here to capture video on the roadside. The proposed method increases the accuracy rate in the detection process.

Ji et al. [14,15] proposed a summarization method based on a multi-video by using archetypal analysis on a multi-modal weighted method. To create WAA weight, the multi-modal graph is used which is done based on the query. A multi-modal graph is used to fuse the information which is generated such as the tags, frames, and video clips for the prediction process. The proposed method outperformed the traditional summarization method by increasing the accuracy rate. The proposed method introduced a sparse coding framework for video summarization using query-aware. The proposed framework uses web images for identifying the exact information of the events. Unsupervised multi-graph fusion is used here to find out the keyframes which are available in the database based on the priority of the queries.

Elharrouss et al. [16] proposed multiple human action detection methods for the recognition and summarization process. Human activities are analyzed and generated into sequences to form a dataset. Then the sequence is divided into shots for the detection process. The histogram of oriented gradient (HOG) is framed based on the frames which are generated by based on the given video clips. The proposed method increases the efficiency and accuracy of the recognition and summarization process.

Zhang et al. [17] proposed a method that uses the key contents of the frames from the given video. A discriminator is used to find out the keyframes for the summarization process. The proposed approach increases the efficiency and accuracy rate.

Yang et al. [18] proposed a new framework using a deep neural network to leverage the benefits generated by the systems. It uses LSTM to represent the priority of the queries which are captured by the network. The proposed method increases the accuracy rate.

Gao et al. [19] proposed a key framework for the video summarization process of surveillance videos. Videos are sequenced based on the overlapping maps features. The clustering approach is used to finalize the key frames and generate an accurate set of frames for further use. The proposed method increases the performance and effectiveness of the system.

Lei et al. [20] have introduced a video summarization model using action parsing driven by a reinforcement algorithm. Action parsing is used to divide the videos into a sequenced part which is used in the final stage. The proposed system deals with recurrent neural networks used in the summarization process which selects the frames based on the actions and activities. The proposed method increases the accuracy rate and classification rate of key frames.

Ji et al. [21] have proposed a new video summarization method by combining a deep attentive ad semantic preserving approach. The Huber loss approach is used to replace the error loss which is occurred during summarization. A deep learning approach is used to ensure the security and safety of the keyframes. The proposed framework increases the performance and robustness of the system.

The proposed method is designed for mitigating the inconsistencies in the frame series detection process. In the MWAA process, the graph alignments are based on weights that imbalance the detection due to frame segregation. Contrarily, the sparse representations in the proposed ERA-SS increase the complexity due to multiple superframes. In this process, the computing time is hiked due to frequent switches over. Therefore, these drawbacks increase the difference in pixel representations, resulting in errors.

III. PROPOSED METHOD

The proposed method intakes video inputs for analyzing its sequence and event detection. The input videos are segregated as frames from which distinct features are extracted for analysis and classification. The data from external dataset is used for validating the proposed method. The input is split into different parts for individual processing as presented in the below Fig. 1. Based on the gate assignments for the observed variations are presented for analysis. In Fig. 1, the proposed method's process is illustrated.

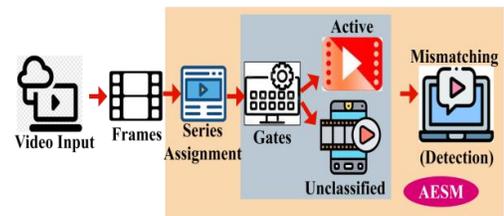


Fig. 1. Proposed Method.

In the series assignment (as in Fig. 1), the mixed heterogeneous features namely contrast (∇) and entropy (t°) are analyzed. First, these two features are extracted from the frames as defined in (1).

$$\left. \begin{aligned} \Delta &= \sum_{i=1, j=1}^n d_{i,j}, \text{diff}(i,j)^2 \\ &\text{and} \\ E^\circ &= \sum_{i=1, j=1}^n -\ln(d_{i,j}) \cdot d_{i,j} \\ &\text{such} \\ &i, j \in \frac{n \times m}{\text{even}} (\text{or}) \frac{m \times n}{\text{uneven}} \end{aligned} \right\} \quad (1)$$

In equation (1) the variables $d_{i,j}$ and $(m \times n)$ represent the pixel density for a frame of size $(m \times n)$. Here i and j balance to m and n for uneven pixel frames or n and n for even pixel frames. The computation $\text{diff}(i,j)$ illustrates the variations between two consecutive pixels. Let T denote the time frame sequence for observing $d_{i,j}$ such that the series assignment is mapped as denoted in (2).

$$\left. \begin{aligned} \forall d_{i,j} \in n \text{ or } m, \Delta &= d_{i,j} dT \\ \text{provided } i * j \text{ and } \nabla &= 1, \text{ if } i = j \\ &\text{and} \\ E^\circ &= \begin{cases} -\ln(d_{i,j}) \forall i * j \\ 0 \forall i * j \end{cases} \\ \text{Therefore} \\ \nabla :: (i,j) \forall T &= \begin{cases} 1 \\ 0, \text{ otherwise} \end{cases} \\ E^\circ :: \text{or } \forall T &= \begin{cases} 1 \\ 0, \text{ otherwise} \end{cases} \end{aligned} \right\} \quad (2)$$

This series assignment as in (3) is used for assigning gates in the learning process. This assignment requires $d_{i,j}$ based assignment in improving the fidelity of summarization. The contrary process of active (sequence) and unclassified is performed. For this purpose, we define current and update gates for state updates. This process is explained in the following subsection.

A. Event Classification

The events are classified by t_o variations in the observed sequences, for which the mapping in equation (2) is used. First, the gates are defined for sequence mapping as in equation (3).

$$\left. \begin{aligned} C &= \tan h \left[\frac{n(i,i)+n(j,j)}{n(i,j)} + \gamma_T \odot \frac{d(i,j)}{n(i,j)} \right] \\ &\text{and} \\ \gamma_T &= \frac{d(i,j)}{n(i,j)} \cdot \frac{n(i,i)+n(j,j)}{n(i,j)} + \int d_{ij} \forall \begin{matrix} i \in n \\ \text{or } j \in m \end{matrix} \end{aligned} \right\} \quad (3)$$

In equation (3), the variables C and γ_T represents the current and update states at time T . Based on the further requirement, the gate states are changed and hence the classifications are performed. In the classification process, the mapping discreteness is observed for detecting active and unclassified series. This detection is performed until the end of the frame. If N is the end of the frame $\forall n, m \in N$, then:

$$\left. \begin{aligned} \Delta_1 &= 1 \\ \Delta_2 &= 1 - \frac{C_1}{n(i,j)} - \frac{C_1}{\gamma_2} \\ &\vdots \\ \Delta_N &= 1 - \frac{C_{N-1}}{n(i,j)} - \frac{C_{N-1}}{\gamma_N} \end{aligned} \right\} \text{for active } N \text{ classification} \quad \left. \begin{aligned} \Delta_1 &= 0 \\ \Delta_2 &= \frac{\gamma_1}{N} + \frac{\gamma_1}{n(i,j)} \times \frac{1}{N} \\ &\vdots \\ \Delta_N &= \frac{\gamma_{N-1}}{N} + \frac{\gamma_{N-1}}{n(i,j)} \times \frac{1}{N} \end{aligned} \right\} \text{for unclassified } N \quad (4)$$

Based on the above classification, the mismatching and detection processes are differentiated. In this process, the E° and ∇ based mismatching for mapped instances as in (2) is performed. The two conditions for ∇ and E° based on T requires multi-feature analysis for a gate assignment. The similarity feature is verified for the stored and acquired features from the mapping as in (3).

$$\left. \begin{aligned} S(x|i, y|j) &= \frac{(2 \times \mu_n \mu_m + k_1)(2 \times \sigma_{nm} + k_2)}{\mu_n^2 + \mu_m^2 + k_1} \frac{(\sigma_n^2 + \sigma_m^2 + k_2)}{\sigma_n^2 + \sigma_m^2 + k_2} \\ &\text{where} \\ \mu &= \frac{1}{n-1} \sum_{i=1}^n (n-i)(m-j) \\ &\text{and} \\ \sigma &= \sqrt{\frac{1}{(n-1)} \sum_{i=1}^n (i-j)^2} = \frac{1}{\sqrt{n-1}} \sum_{i=1}^n (i-j)^2 \forall i \neq j \end{aligned} \right\} \quad (5)$$

This similarity verification is performed for different $(i, j) \in (n, m) \in N$ wherein the mismatch for the above requires an alternate mapping such that C is assigned with a new $n(i, j)$ and γ_T is updated for $(n-1)^{th}$ $d(i, j)$. This means the variations in consecutive pixels are violated in detecting an event. The detection is performed in multiple intervals from 1 to N such that the mismatching $d(i, j)$ are segregated. A contrary part of the abnormal event detection is the synchronization of the $(1 - \frac{S}{N})$ and mapping as in (2) for the different features. In the proposed method, the abnormal events at different T are considered non-cumulative (due to different occurrences). Therefore, the occurrences are synchronized based on $S(x|i, y|j) \forall n$ and m until N is achieved. This is validated for $d(i, j)$ such that the alternating sequences are varied until the end of classification. In Fig.2, the gate assignment and classification processes are illustrated.

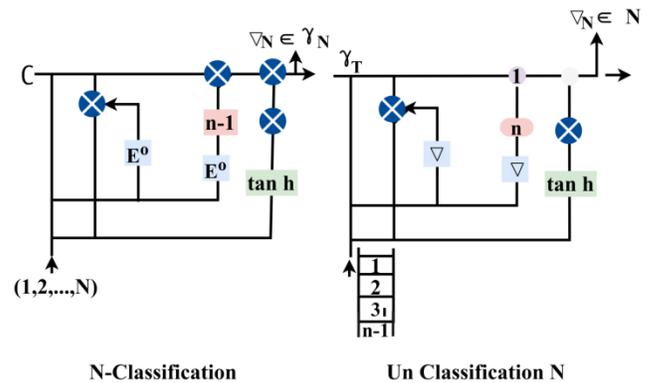


Fig. 2. Classification Process.

In the classification process, as presented in Fig. 2, the "×" and "+" symbols represent the product and sum of the mapping presented in equation (2). First, the product represents the $\nabla \otimes E^\circ \forall T$ in 1 to n (or) 1 to m ; the sum is the joint set of ∇ and E° . For an abnormal event summarization, the $T \notin (\nabla \odot E^\circ)$ is segregated for classifying it as a whole interval. In contrast to the augmenting and mapping processes, the variations are detected for event detection and categorization. Therefore,

$$\left. \begin{aligned}
 & d_{i,j} \in T \notin (\nabla \odot E^\circ) \text{ is expected for } T = 1 \\
 & \quad \text{rather,} \\
 & (i,j) \in n \notin N \forall \nabla \text{ is achieved for } T = 0 \\
 & \quad \text{such that} \\
 & \quad \Delta_N \in n(i,j) \forall C \text{ and} \\
 & \quad \Delta_N \in N \forall \gamma_T \mu \text{ is high}
 \end{aligned} \right\} \quad (6)$$

In equation (6) the abnormal (Post the matching) T is identified for summarization. In the summarization process, the distinct event occurrences are augmented cumulatively. The differences are mitigated without augmenting the Δ_N as classified for $\gamma_T \in N$ and $S(x|i, y|i)$. The summarization process is described below.

B. Summarization Process

In the summarization process, the γ_T that encloses both $\Delta_N \in \gamma_T$ and $s(x|i, y|j)$ (failing) conditions are augmented based on T . This is either discrete/ sequential depending on multiple updates as in E° and ∇ . The process requires unidentified $d(i,j)$ post the gate allocation for maximizing event aggregation. If the event is observed in $T \forall \Delta_N \in \gamma_T$ interrupts, then.

$$\left. \begin{aligned}
 & t_{i=1}^T = |\Delta_N^2 - \max\{C, E^\circ\}|_{i=1 \text{ to } T} \text{ and} \\
 & \quad \text{(or)} \\
 & t_{i=T-t}^T = \left| \frac{\Delta_N}{N} - \min\{E^\circ, \gamma_T\} \right| \forall i = T - t \text{ to } t
 \end{aligned} \right\} \quad (7)$$

In equation (7), the augmenting events classified for $i = t$ and $i - (T - t)$ are identified. Depending on the $\max\{C, E^\circ\}$ and $\min\{E^\circ, \gamma_T\}$ the abnormal classifications is grouped. This is required for projecting $d(i,j) \in T$ and hence the deviations are identified. The proposed method performs a cumulative augmentation of the above observation post $(\nabla \odot E^\circ)$ assessment and hence the summary is an allocation of $d(i,j) \in$ distinct T and $(T - t)$. This is non-recurrent and hence new frames in T and $(T - t)$ (intermediate) are identified without false rates.

IV. EXPERIMENTAL RESULTS AND DISCUSSION

This section discusses the proposed method's performance assessment using MATLAB simulations. The dataset from UCSD [22] is used for validating the proposed method's performance for accuracy, false rate, analysis time, SSIM, and F1-Score. The inputs are classified based on the available objects; the objects are used as in the dataset labeling. Depending on 4 textural features, the classification is performed; the pixel un-matching inputs are alone mitigated. In this comparative analysis, the identified objects and state updates are varied for the proposed and existing MVS-MWAA [14] and ERA-SS [12] methods. The data set provides

multiple video frames observed at 30fps in 800x480 pixel resolution. A total of 9390 frames are observed in this dataset.

A. Accuracy

In Fig. 3, the accuracy for different objects and state updates is analyzed. The proposed method maximized accuracy by improving the γ_T and E° detection in T and C and Δ_N detection $\forall (T - t)$. This is non-recurrent based on the available $n(i,j)$ in multiple T such that accuracy is maximized. Another detection is the $s(x|i, y|j) \in N$ where multiple pixels with the observed features are validated. This is consistent for different objects identified in the frames wherein accuracy is high.

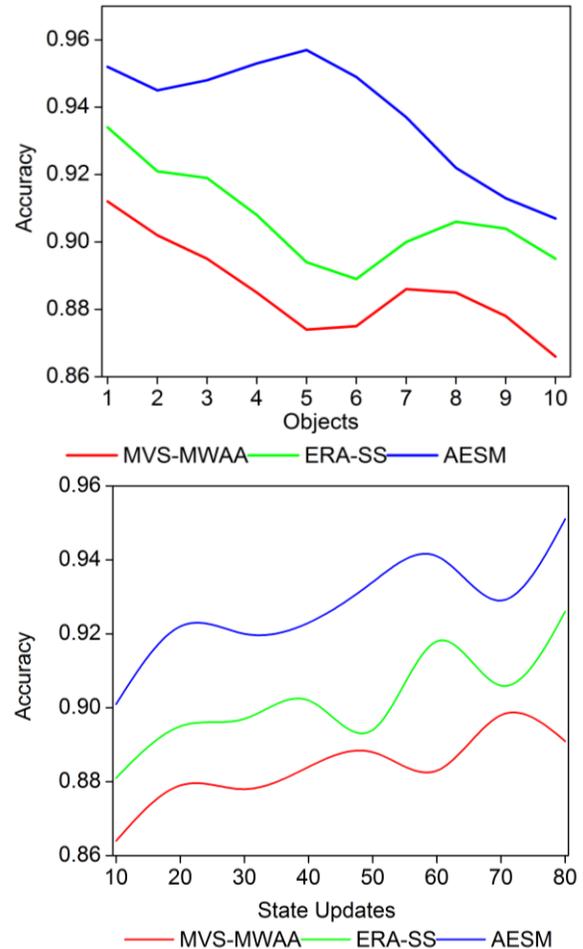


Fig. 3. Accuracy Analysis.

B. False Rate

The augmentation of $(\Delta \odot E^\circ)$ and $(\nabla \cup E^\circ)$ relies on multiple factors of C and E° such that no error arises. The proposed requires Δ_N classification based on C and γ_T such that in t interference is avoided. In distinct instances, interferences are modeled independently.

The gate updates are non-linear $\forall S(x|i, y|j)$ for unclassified N such that C is high. In this process, the unclassified instances are reduced which achieves less False rate (refer to Fig. 4).

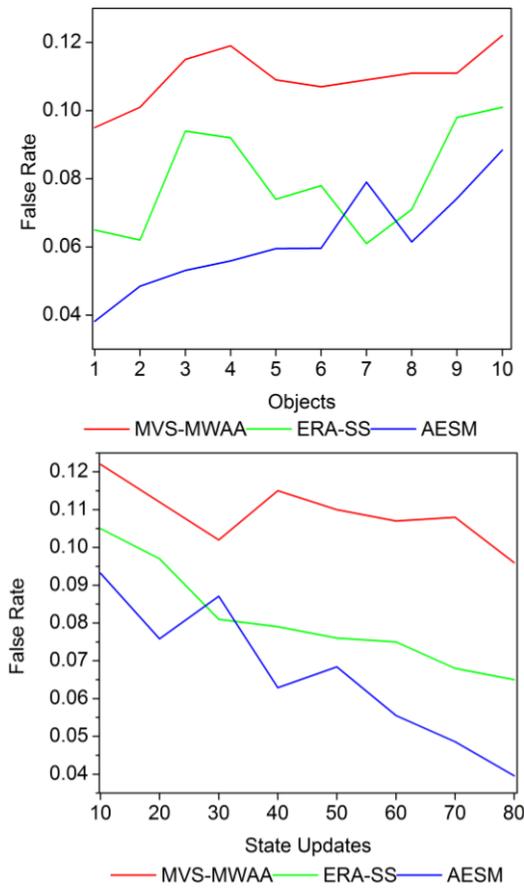


Fig. 4. False Rate Analysis.

C. Analysis Time

The proposed method achieves less analysis time as the proposed method classifies C and $\gamma_T \forall d(i, j)$. In the active classification and (μ, σ) estimation, independent assessments are performed.

These are validated based on the mapping and hence $(T - t)$ and T as independent rather than cumulative and joint analysis. Therefore, the proposed method achieves less analysis time for identified objects and state updates (Fig. 5).

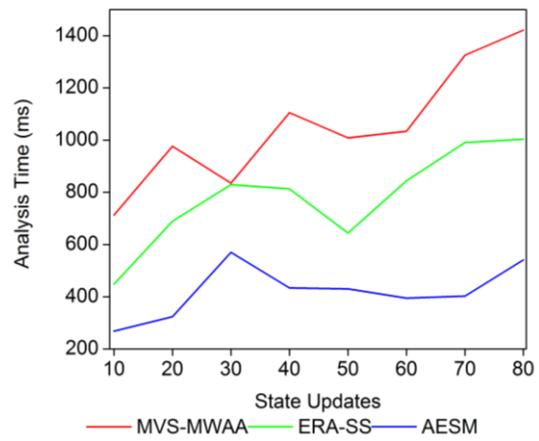
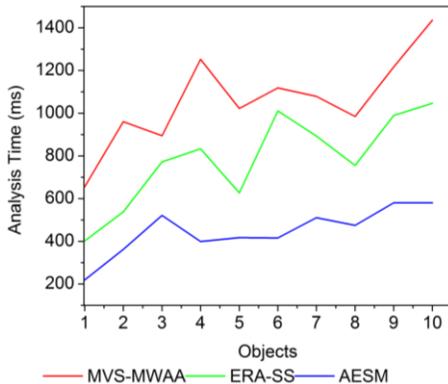


Fig. 5. Analysis Time.

D. SSIM

In Fig. 6, the SSIM for the proposed method is compared for different objects and state updates.

In multiple state updates, the classification is performed under different N . These classifications are performed for $\Delta_N \in \gamma_N$ and $\Delta_N \in N$ for detecting multiple SSIM for $d(i, j)$ such that \forall is achieved in different $(T - t)$.

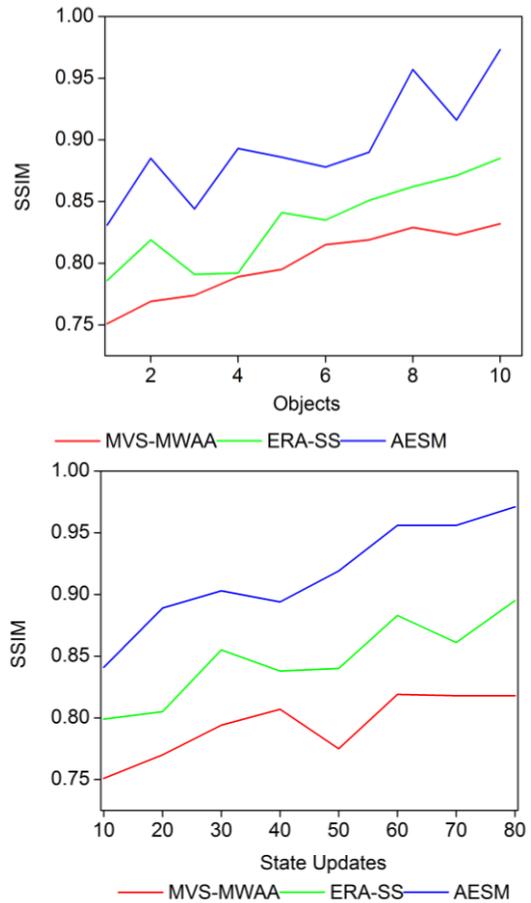


Fig. 6. SSIM Analysis.

This is unanimously observed for distinct intervals and objects where γ_T is less, maximizing SSIM.

E. F1-Score

For any density of objects and state updates, the F1-score is high for the proposed method (Fig. 7). The proposed method achieves a high F1score by mitigating ($\nabla \cup E^\circ$) instances. This is performed based on the classification and mapping of distinct $d(i, j)$. In the classification process, ($T - t$) and T instances are distinguished for maximizing the F1-score, preventing false rate. The comparative analysis results are tabulated in Tables I and II for different objects and state updates.

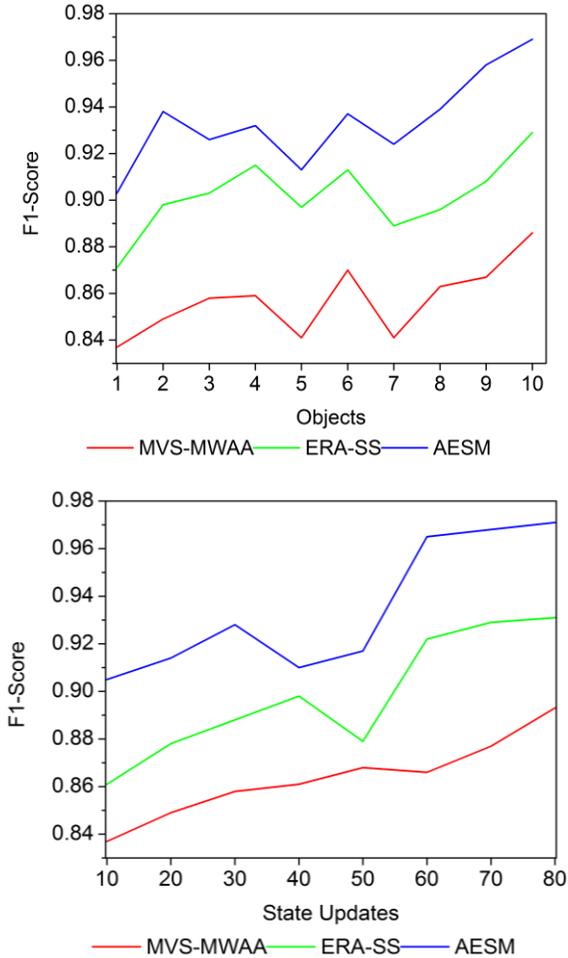


Fig. 7. F1-Score Analysis.

TABLE I. COMPARATIVE ANALYSIS RESULTS FOR OBJECTS

Metrics	MVS-MWAA	ERA-SS	AESM	Findings
Accuracy	0.866	0.895	0.907	8.83% High
False Rate	0.122	0.101	0.0884	7.7% Less
Analysis Time (ms)	1436.06	1046.51	580.226	8.88% Less
SSIM	0.832	0.885	0.973	11.45% High
F1-Score	0.886	0.929	0.969	12.3% High

TABLE II. COMPARATIVE ANALYSIS RESULTS FOR STATE UPDATES

Metrics	MVS-MWAA	ERA-SS	AESM	Findings
Accuracy	0.891	0.926	0.951	7.08% High
False Rate	0.096	0.065	0.0396	6.82% Less
Analysis Time (ms)	1420.73	1003.2	541.227	9.22% Less
SSIM	0.818	0.895	0.971	11.55% High
F1-Score	0.893	0.931	0.971	11.8% High

The significance of the proposed method is adaptable for varying objects and computations (state updates). In a video processing, the variations due to objects and computations are practically addressed using this proposed method. The variations are suppressed using the gate assignment and hence the accuracy, SSIM, and F1-Score are improved.

V. CONCLUSION

This article discussed an additive event summarization method for reducing the inconsistencies in video event summarization. The classification events are identified using different state assignments through gated recurrent units. The recurrent unit identifies unclassified and active frames for preventing false rates in event extraction. This classification is performed based on the heterogeneous features over the varying pixel densities over different sequences. From the analyzed sequences, the abnormal feature exhibiting pixels are segregated for providing a summarized output. For the different objects classified, the proposed method achieves 8.83% high accuracy, 11.45% high SSIM, 12.3% high F1-score, 7.7% less false rate, and 8.88% less analysis time. Though the proposed method is reliable in summarizing event related to abnormal occurrences the varying textural features result in pixel errors. Therefore, a spatiotemporal feature classification pre-processing is planned to be integrated in the future work.

REFERENCES

- [1] Wang, H., Feng, J., Sun, L., An, K., Liu, G., Wen, X., ...& Chai, H. (2020). Abnormal Trajectory Detection Based on Geospatial Consistent Modeling. *IEEE Access*, 8, 184633-184643.
- [2] Wang, X., Song, H., & Cui, H. (2018). Pedestrian abnormal event detection based on multi-feature fusion in traffic video. *Optik*, 154, 22-32.
- [3] Wu, J., Xu, H., Zheng, Y., & Tian, Z. (2018). A novel method of vehicle-pedestrian near-crash identification with roadside LiDAR data. *Accident Analysis & Prevention*, 121, 238-249.
- [4] Alhussain, T. (2021). Density-scaling traffic management for autonomous vehicle environment—predictive learning-based technique. *Soft Computing*, 1-15.
- [5] Jiang, Z., Liu, Y., Fan, X., Wang, C., Li, J., & Chen, L. (2020). Understanding urban structures and crowd dynamics leveraging large-scale vehicle mobility data. *Frontiers of Computer Science*, 14(5), 1-12.
- [6] Yang, L., & Yang, N. (2019). An integrated event summarization approach for complex system management. *IEEE Transactions on Network and Service Management*, 16(2), 550-562.
- [7] Muhammad, K., Hussain, T., Del Ser, J., Palade, V., & De Albuquerque, V. H. C. (2019). DeepReS: A deep learning-based video summarization strategy for resource-constrained industrial surveillance scenarios. *IEEE Transactions on Industrial Informatics*, 16(9), 5938-5947.

- [8] Fei, M., Jiang, W., & Mao, W. (2021). Learning user interest with improved triplet deep ranking and web-image priors for topic-related video summarization. *Expert Systems with Applications*, 166, 114036.
- [9] Lan, L., & Ye, C. (2021). Recurrent generative adversarial networks for unsupervised WCE video summarization. *Knowledge-Based Systems*, 222, 106971.
- [10] Zhang, J., Shi, Y., Jing, P., Liu, J., & Su, Y. (2019). A structure-transfer-driven temporal subspace clustering for video summarization. *Multimedia Tools and Applications*, 78(17), 24123-24145.
- [11] Yang, M., Qu, Q., Shen, Y., Zhao, Z., Chen, X., & Li, C. (2020). An Effective Hybrid Learning Model for Real-Time Event Summarization. *IEEE Transactions on Neural Networks and Learning Systems*.
- [12] Wan, S., Xu, X., Wang, T., & Gu, Z. (2020). An intelligent video analysis method for abnormal event detection in intelligent transportation systems. *IEEE Transactions on Intelligent Transportation Systems*.
- [13] Thomas, S. S., Gupta, S., & Subramanian, V. K. (2017). Event detection on roads using perceptual video summarization. *IEEE Transactions on Intelligent Transportation Systems*, 19(9), 2944-2954.
- [14] Ji, Z., Zhang, Y., Pang, Y., Li, X., & Pan, J. (2019). Multi-video summarization with query-dependent weighted archetypal analysis. *Neurocomputing*, 332, 406-416.
- [15] Ji, Z., Ma, Y., Pang, Y., & Li, X. (2019). Query-aware sparse coding for web multi-video summarization. *Information Sciences*, 478, 152-166.
- [16] Elharrouss, O., Almaadeed, N., Al-Maadeed, S., Bouridane, A., & Beghdadi, A. (2021). A combined multiple action recognition and summarization for surveillance video sequences. *Applied Intelligence*, 51(2), 690-712.
- [17] Zhang, Y., Kampffmeyer, M., Liang, X., Zhang, D., Tan, M., & Xing, E. P. (2019). Dilated temporal relational adversarial network for generic video summarization. *Multimedia Tools and Applications*, 78(24), 35237-35261.
- [18] Yang, M., Tu, W., Qu, Q., Lei, K., Chen, X., Zhu, J., & Shen, Y. (2019). MARES: multitask learning algorithm for Web-scale real-time event summarization. *World Wide Web*, 22(2), 499-515.
- [19] Gao, Z., Lu, G., Lyu, C., & Yan, P. (2018). Key-frame selection for automatic summarization of surveillance videos: a method of multiple change-point detection. *Machine Vision and Applications*, 29(7), 1101-1117.
- [20] Lei, J., Luan, Q., Song, X., Liu, X., Tao, D., & Song, M. (2018). Action parsing-driven video summarization based on reinforcement learning. *IEEE Transactions on Circuits and Systems for Video Technology*, 29(7), 2126-2137.
- [21] Ji, Z., Jiao, F., Pang, Y., & Shao, L. (2020). Deep attentive and semantic preserving video summarization. *Neurocomputing*, 405, 200-207.
- [22] <https://gram.web.uah.es/data/datasets/rtm/index.html>

Relational Deep Learning Detection with Multi-Sequence Representation for Insider Threats

Abdullah Alshehri

Department of Information Technology
Faculty of Computer Science and Information Technology
Al Baha University
Alaqlq 65779-7738, Saudi Arabia

Abstract—Insider threats are typically more challenging to be detected since security protocols struggle to recognize the anomaly behavior of privileged users in the network. Intuitively, an insider threat detection model depends on analyzing the audit data, representing trusted users' activity streams, on recognizing malicious behaviors. However, the audit data is high dimensional data in that it presents n dependent streams of activities where it establishes a complex feature extraction. In this context, the dependent streams represent user activities where each activity is represented by an ordered set of real variables that pertain to a specific occurrence, such as log-in records. As a result, multiple actions can be represented simultaneously, with one or more values being recorded at each timestamp. Moreover, the relations between dependent streams are typically neglected while detecting the anomaly behavior. Ideally, relation learning is commonly considered to recognize occurrence patterns in streaming data. Thus, the latent relations are thought to have insight for the accurate detection of anomaly behavior concerning insider threats. This study introduces a novel model to detect insider threats by representing audit data as multivariate time series to explicitly learn the existing inter-relations between activity streams using a Recurrent Neural Network (RNN). The model considers learning the latent relationships to effectively extract features for modeling the behavior profile where anomaly behavior can be detected accurately. The evaluation, using the CERT dataset has shown that the proposed model outperforms the comparator approaches to insider threats detection with AUC of 0.99.

Keywords—Insider threats; machine learning; recurrent neural network; user behavior analytic

I. INTRODUCTION

With the rapid advancement and growth in networking technology, cyber threats have become a significant issue for numerous companies and organizations worldwide [1]. A cyber threat can mainly be realized by breaching network security. Ideally, the primary option for malicious intent to breach network security is by using a malware [2]. In this context, the malware contravenes the secured network by an external malicious component such as rootkits, Trojan horses, viruses, and worms [3]. Thus, the ideal solution to secure the network from such external threats is by proposing a perimeter defense, e.g., firewalls, antivirus software, and intrusion prevention/detection systems.

However, a cyber attack can be triggered from an internal source in the network; it is well known as an insider threat [4]. A typical form of an insider threat is that the legitimate user may conduct harmful work at the network, such as

leaking, altering, or disrupting sensitive data. Thus, an insider threat (malicious) is typically realized as an abnormal action, or behavior, in the network flows that is performed by the legitimate user [5]. Fig. ?? shows a conceptual illustration of the internal and external attacks intuition which can affect such a network in cyber space. It can be observed that the external attack is transparent to be prevented/detected by the perimeter defense protocols. On the other hand, the insider attack is commonly deceptive as the perimeter defense can hardly detect attacks conceived from inside the network. Therefore, insider attacks pose a critical challenge in the cyber security domain where the demand to propose practical solutions to detect insider threats remains a desirable solution for a tremendous number of organizations in the market [6].

To detect insider threats, the typical solution relies on developing systems that are capable of analyzing the user's behavior to discover anomalies in the network [7]. The idea is to observe the user's daily activities and tasks where these activities yield frequent network usage patterns. Ideally, the regular activities can underline insightful patterns to map a typical behavior for the legitimate user. In this context, the ideal method to analyze the user's behavior is accomplished using Machine Learning (ML)-based approaches, (see for instance [8], [9], [10], [11]). ML approaches, such as Hidden Markov Model (HMM) and Support Vector Machine (SVM), have been utilized to detect insider threats via modeling the behavioral profile from the audit data (daily activities) such as the log events. Typically, these ML approaches, which are well known as shallow learning methods [12], are subject to attentive feature engineering to model the behavior profile accurately. The reason is that the audit data is composed of a large volume of unstructured, high-dimensional, and sparsity instances, which makes extracting features a non-trivial task. The traditional approaches have modeled the user's behavior by aggregated data consisting of the user's activities within a single day. However, missing some features can cause unpredictable behavior where it imposes unbalanced detection alarms; for example, it can increase the false alarms in the system. Deep Learning (DL), a subset of ML approaches, has been employed to address the drawback mentioned earlier for insider threats detection [13], [4], [14], [15]. DL provides the advantage that features can be represented immediately from unstructured data [16]. Moreover, it has the advantage that features can be extracted sequentially to reduce missing temporal feature learning, a way that is not applicable in the case of shallow learning methods.

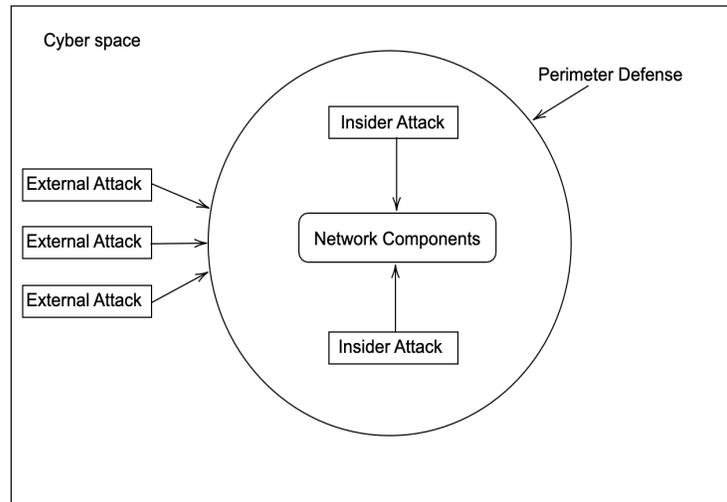


Fig. 1. A Conceptual Illustration Shows the Intuition of External and Internal Attacks in Cyber Space Networks.

Nevertheless, insider threats detection approaches neglect considering the temporal representation of multivariate sequences to model the user's behavior, where this limitation has established a downside to developing an accurate detection model. Intuitively, the user activities are temporally recorded as sequences of dependent variables, i.e., at each timestamp, one or more variables, such as user id and log-in/off time, would be recorded simultaneously. Therefore, incorporating more extensive dependent variables can increase the chance of better modeling the user's behavior. It is worth mentioning that increasing data volume results in improving the learning accuracy as per Bonferonni's principle [17]. This, in turn, brings the motivation to structure the entire audit data as temporal sequences, i.e., multivariate time series representation, which thought it is fruitful to map all possible behavior patterns of the user. Moreover, as the audit data consist of multi-sequential actions, the relations between these sequences are not well considered in previous studies. Thus, the conjecture is that the sequences related to one user would underline strong relations to describe unique user-related patterns used for insider threat detection.

This study proposes a novel model that utilizes DL to learn the user's behavior for insider threats detection using a multivariate representation of audit data. The model represents the user activities as a set of dependent sequences in the temporal domain to where the hidden relations are extracted and learned. In concise, each activity is represented as an ordered set of actual values that refers to some event, such as log-in records. Thus several activities can be represented simultaneously such that one or more values are recorded at each timestamp, i.e. the user activities are represented as multivariate time series streams such that at each time tick, n values are recorded temporally. We then use Recurrent Neural Network (RNN) to learn the existing latent temporal relationships between sequences to map the hidden patterns. Thereafter, the model serves to extract features where the recognized behavior is classified as normal or anomaly, which

indicates a possible insider threat.

The contributions of the proposed work can be summarized as follows:

i) The study has presented a novel method to model the user behavior in a multivariate time series structure. More specifically, the user behavior is represented using all sequences of events that denote the user activities. The intuition is that time series frequencies would present accurate, readable user behavior patterns because they incorporate exclusive features, not only aggregated features.

iii) The developed model has considered learning and extracting the relations between the multivariate sequences to extract patterns in training data. The existing sequences hold latent relations that can be extracted for accurate behavior modeling, leading to the accurate prediction of anomaly behavior.

iii) The proposed model has applied deep learning to extract features from inter-correlation streams that represent audit data of user activities. The importance is that user activities are stochastic and unstructured, so deep learning is ideal for extracting features from unstructured data.

The remainder of this paper is structured as follows. Section II presents an overview of the related work. This is followed by Section III where the proposed model has been introduced. In Section IV, we demonstrate the evaluation and results of the proposed model. Section V provides conclusions and future directions where this study can be extended.

II. RELATED WORK

ML has been increasingly used for cyber threats detection throughout the previous decade [18]. Generally speaking, ML has shown to be advantageous in the identification and classification of anomalous occurrences in the network streams [19], [20]. Insider threat is a well-known type of cyber attack [21] that has received considerable work using ML approaches. The insider attack detection model heavily depends on the user's

daily activities where the behavior is recognized. Typically, the obtained data is unstructured and complex due to the diversity of the user's activities on the network. Thus, modeling the user's behavior is a relatively intricate task. In the literature, most ML approaches for insider threat detection are data-driven in that the user activities are aggregated to underline representative features where the behavior profile is modeled. In this context, insider threat detection would be proposed based on different examples of data instances where the detection model is handled as an anomaly detection problem. Accordingly, various ML methods have been developed for insider threat detection. For example, SVM is used as a one-class detection method for insider attack [22]. The study presented in [9] had proposed an HMM model to map the typical user behavior based on weekly activities. The insider attack is detected by computing a deviation score between sequences; a low probability score could indicate a probable attack. In [8] a set of supervised and unsupervised ML approaches have been evaluated for insider attack detection. The study had used Self Organization Map (SOM), HMM, and Decision Tree to model malicious behavior for anomaly detection. The features were extracted into two categories, including numerical and sequential features. Whatever category was being employed, the features were aggregated to a set of weekly representative instances. Thus the detection model was complex due to the need for extensive feature engineering.

DL methods have been proposed to tackle the issue of requiring feature extraction and handling the large volume of features to be learned in the detection model for insider attacks data. In [5] a comprehensive survey has introduced the state-of-the-art of DL with insider threat detection. In this context, a number of DL applications have been recruited such as deep feed-forward neural network [23], [24], recurrent neural network (RNN) [14], [25], conventional neural network[26], and graph neural network. [13], [27]. Due to the complexity of data structure, the majority of DL approaches have focused on representing sub-sequences, such as one-day activities, for detection granularity. In practice, each session is a subsequence that denotes a series of activities, i.e. "log-in" and "log-off" events. Whenever a subsequence contains malicious activity, the subsequence will be designated as a malicious subsequence where a possible attack could occur. Therefore, detecting abnormal actions is difficult due to the limited information (features) that can be leveraged. Moreover, the relation extraction between sequences is ignored; however, the extraction of latent relations can bring insight for better modeling of user's behavior.

This study addresses the above-mentioned drawbacks by representing all activities (sequences) as multivariate time series streams where the relations between streams are also considered for building the behavior profile. The study also endeavors to leverage the advantage of using RNN for effective feature extraction from the temporal/sequential data. Recall that RNN has shown effective feature learning of the sequential data for anomaly detection [28], [29], [30].

III. PROPOSED MODEL

This section elaborates on the proposed insider threat detection model. Fig. ?? illustrates an overview of the model flows. As it can be seen from the figure that the model

operates in several consecutive steps, including i) user activity representation, ii) sequence activity embedding, iii) latent relations learning, iv) feature learning, and v) anomaly detection. The following subsections give further detail of each step as follows.

A. User Activity Representation

The primary step in the proposed model is to represent the user's activities as multivariate time series. As noted in the introduction to this paper, the user's activities can be structured as a series of temporal activities recorded each tick of time. Several data points (characteristics values) should be recorded at each timestamp, such as log-in/off information, user's id, and HTTP data.

More formally, given a series of a single activity \mathcal{A}_1 it consists of a sequence of an ordered n data points such that $\mathcal{A}_1 = \{p_1^1, p_1^2, \dots, p_1^n\}$, as for a given point p it maps an encoded real value of an event. The entire activities are expressed as a whole series \mathcal{S} which represent set of m activities, i.e. $\mathcal{S} = (\mathcal{A}_1, \mathcal{A}_2, \dots, \mathcal{A}_m) = (\mathcal{A}^1, \mathcal{A}^2, \dots, \mathcal{A}^m)^\top \in \mathbb{R}^{m \times n}$, where $n \in \mathbb{N}$ is the length of time series, i.e. the number of data points in each single activity. Intuitively, \mathcal{S} is structured as a matrix consisting the entire user's activities whose samples are denoted as $p_i^{(t)}$ ($i = 1, \dots, m; t = 1, \dots, n$).

$$\mathcal{S} = \begin{bmatrix} p_1^1, p_1^2, \dots, p_1^n \\ p_2^1, p_2^2, \dots, p_2^n \\ \dots \\ p_m^1, p_m^2, \dots, p_m^n \end{bmatrix}$$

B. Sequence Activity Embedding

The represented activity sequences have a wide range of features that can be related in diverse ways. For example, if we consider the log-in and user id sequences for two different users, the sequences for the same user likely have strong relations. Thus, the idea is to portray each sequence in a flexible fashion that captures the various characteristics that underpin its behavior in a multidimensional manner. To this end, each activity sequence \mathcal{A} has been encoded as an embedding vector \vec{v} such that $\vec{v}_i \in \mathbb{R}^d$, for $i \in \{1, \dots, m\}$. Note that the encoded embedding sequences are randomly initialized before being trained with the remainder of the model. Moreover, sequences with comparable embedding values should have a strong inclination to be connected since similar embedding sequences indicate similar activities.

C. Latent Relations Learning

The subsequent step is to learn relations between the embedded vectors. The optimal method to conduct so is by using direct graph architecture. In this context, given embedded vectors $\mathcal{V} = \{\vec{v}_1, \dots, \vec{v}_{|\mathcal{V}|}\}$, they are mapped to graph structure $\mathcal{V} \mapsto (\mathcal{N}, \xi)$ with nodes and edges; where nodes denote embedded vectors, such that $\vec{v}_i \in \mathcal{N}$ for $i = \{1, \dots, |\mathcal{N}|\}$, and edges represent pairs $\varepsilon_i \in \xi \mapsto \varepsilon_i = \langle \vec{v}_i, \vec{v}_j \rangle \in \mathcal{N} \times \mathcal{N}$. In this study, we implement direct graph representation for latent relations learning as direct edges $\vec{v} \mapsto \vec{v}_i$ between nodes because the dependency patterns between vectors do not have to be symmetric. Thus, the mapped edges between vectors represent relative dependant relationship in that the first vector

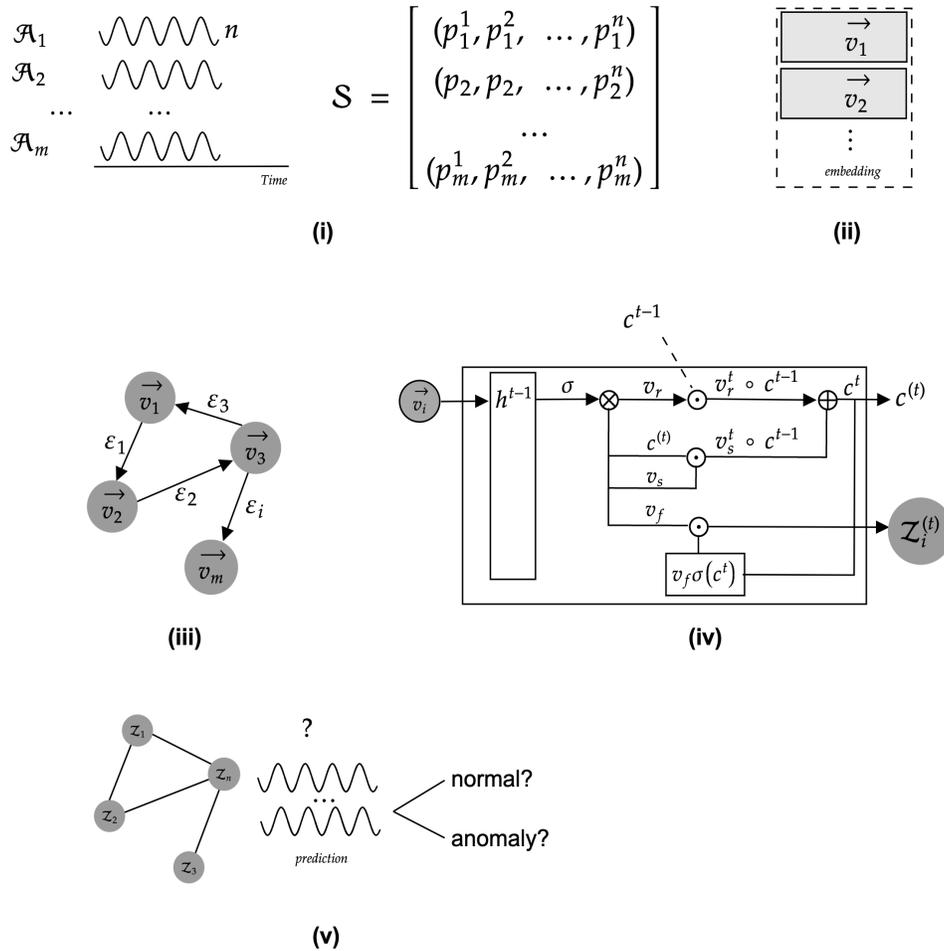


Fig. 2. Illustration of the Proposed Insider Threat Prediction Model.

is used to model the behavior of the second vector. Recall that given a node vector \vec{v} is denoted by $h_{\vec{v}} \in \mathbb{R}^d$. Each node has given a label $\ell_{\vec{v}} \in \{1, \dots, \mathcal{L}_N\}$ that indicates the activity type, e.g. the log-in activities, where the edge has also been given label such that $\ell_{\epsilon} \in \{1, \dots, \mathcal{L}_{\epsilon}\}$ for each given ϵ_i .

Note that when an edge connects two vectors, it means the first vector is utilized to underline the behavior of the second vector. The dependency is represented between vectors as a set of candidate relations \mathcal{R}_i for each vector such that $\mathcal{R}_i \subseteq \{1, \dots, m\} \setminus \{\vec{v}_i\}$. The selection of which dependencies related to \vec{v}_i is conducted by the search for the most similar candidate. To this end, we compute the similarity θ between ϵ_i edge node and the embeddings of its candidate relation $j \in \mathcal{R}_i$ using dot product measure. Equation 1 shows θ similarly measurement.

$$\theta(\epsilon_i, \epsilon_j) = \frac{\vec{v}_i \cdot \vec{v}_j}{\|\vec{v}_i\| \cdot \|\vec{v}_j\|} \quad \text{for } j \in \mathcal{R}_i \quad (1)$$

D. Feature Learning

Having represented relations between embedded vectors (graph nodes), the next step is concerned with extracting and learning features. The idea is to establish an abstracted feature space using RNN to fuse a node's information with its neighbors. In the proposed model, feature extraction includes

the vector embedding \vec{v}_i , which describes the various behaviors of various vector kinds. Nevertheless, feature extraction has been accomplished using Long-Short Term Memory (LSTM), a well-known set of RNN architecture. The main benefit of adopting an LSTM unit is that the cell state averages activities over time, which helps to avoid disappearing gradients and better capture long-term time series relationships.

At each tick of time, an individual entry node \vec{v} will map a hidden layer $h^{(t)}$. Each hidden unit has a memory cell $c^{(t)}$ to obtain long-term dependencies. The intuition is that $c^{(t)}$ serves to remember the effect of the prior input layer. In the proposed model, the mapping function use three non-linear gates to manage the access to $c^{(t)}$ cell as follows: i) remember vector v_r , save vector v_s , and focus vector v_f . The following equations express the mathematical notation of gated vector v_r, v_s, v_f respectively (2, 3, 4), memory cell control $c^{(t)}$ (5), and mapping function $h^{(t)}$ (6).

$$v_r = \sigma(W_r[h^{(t-1)}; v^{(t)}] + \epsilon_r) \quad (2)$$

$$v_s = \sigma(W_s[h^{(t-1)}; v^{(t)}] + \epsilon_s) \quad (3)$$

$$v_f = \sigma(W_f[h^{(t-1)}; v^{(t)}] + \epsilon_f) \quad (4)$$

$$c^{(t)} = v_r \odot c^{(t-1)} + v_s \odot \sigma(W_c[h^{(t-1)}; \vec{v}^{(t)}] + \epsilon_c) \quad (5)$$

$$h^{(t)} = v_f \odot \sigma(c^{(t)}) \quad (6)$$

where $[h^{(t-1)}; v^{(t)}] \in \mathbb{R}^d$ is the sum of the prior hidden state $h^{(t-1)}$ and the current vector $v^{(t)}$ along with some bias ϵ , \odot is the element-wise multiplication, and σ is the non-linear Rectified Linear Unite (ReLU) activation function [31]. Here, the resulted output is an aggregated representation \mathcal{Z}_i of hidden layers at time (t) from such input node \vec{v}_i .

$$\mathcal{Z}_i^{(t)} = \sigma(h^{(t)}, \vec{v}_i^{(t)}) \quad (7)$$

where $\vec{v}_i^{(t)}$ is the input for a given node at time t , with ReLU activation σ .

E. Anomaly Detection

When the relations are learned as per the previous subsection, the next step is to determine how the anomaly behavior can be detected accordingly. The idea is to determine how such an unseen behavior deviates from learned relations for each user. Thus, the proposed model attempts to calculate a similarity score Φ between the observed behavior of the user, that has resulted from the learned relations, and the new abstracted stream $\hat{\mathcal{Z}}$:

$$\Phi_i^{(t)} = |\mathcal{Z}_i^{(t)} - \hat{\mathcal{Z}}_i^{(t)}| \quad (8)$$

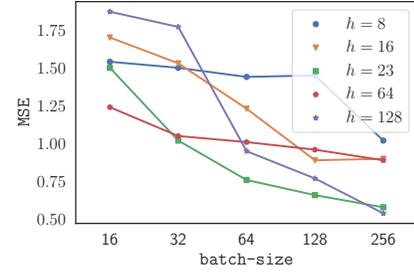
To assure a robustness calculation of the similarity score, we normalize the score for each input node using the median u of the difference between 1st and 3rd quartiles of distribution α .

$$\varphi_i^{(t)} = \frac{\Phi_i^{(t)} - u_i}{\alpha_i} \quad (9)$$

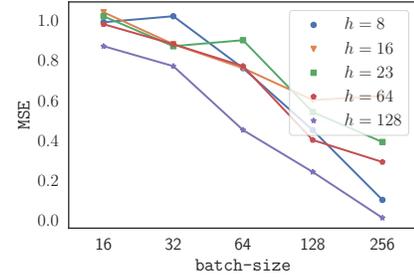
Recall that the use of inter-quartile range shows an effectiveness calculation of the distribution's spread for anomaly detection in stream data [32]. Then, the anomaly score Anom_{sc} is the max value of φ that is computed at time (t) as follows:

$$\text{Anom}_{sc}^{(t)} = \max(\varphi_i^{(t)}) \quad (10)$$

Hence, the stream is classified as either normal or anomaly activity at some fixed threshold. The user can configure the threshold value; however, in the evaluation of this study, the value is set to max over the validation data. Thus, the stream is labeled as an anomaly whenever the similarity score exceeds the max $\text{Anom}_{sc}^{(t)}$ value.



(a)



(b)

Fig. 3. The MSE Performance over the Grid Search Concerning different Parameters: (a) Considers the Grid Search with epochs = 50, and (b) Considers the Grid Search with epochs = 100

IV. PERFORMANCE EVALUATION

This section provides the evaluation of the proposed model to determine its efficacy. The main objective of the evaluation is to show the effectiveness of detecting insider anomalous in network flows based on relation-based learning with LSTM. Moreover, the evaluation examines the model's performance under different LSTM parameters tuning to determine how they would affect the detection accuracy. Finally, the evaluation shows how well the performance of the proposed model compared to baseline methods of detecting insider attacks.

A. Evaluation Settings

1) *Dataset*: The evaluation experiments have been conducted over CERT dataset [33]. CERT is a public released insider threat dataset. It consists of activities data for more than 1k users, with 32 million events (log lines) generated over 502 days. There are around 7k log lines representing anomaly actions among the total recorded activities; these logs were manually placed into the data records by specialists. The data pertaining to log-in, log-off, device, and HTTP is stored in the logline. Each user action is parsed into a vector in the experiment, including the id, date, user, computer, and activity type as multivariate time series sequences. The dataset has been splitted to training (70%) and testing (30%) for all conducted experiments.

2) *Metrics*: Confusion Matrix (CM) has been used to compute a number of metrics to evaluate the model's performance. Recall that CM shows the classification performance of the proposed model under different parameters settings and with compared to other comparators baselines. Thus, CM is used to measure: i) F1-score ($F1 = \text{Equation 11}$), Precision ($Pre =$

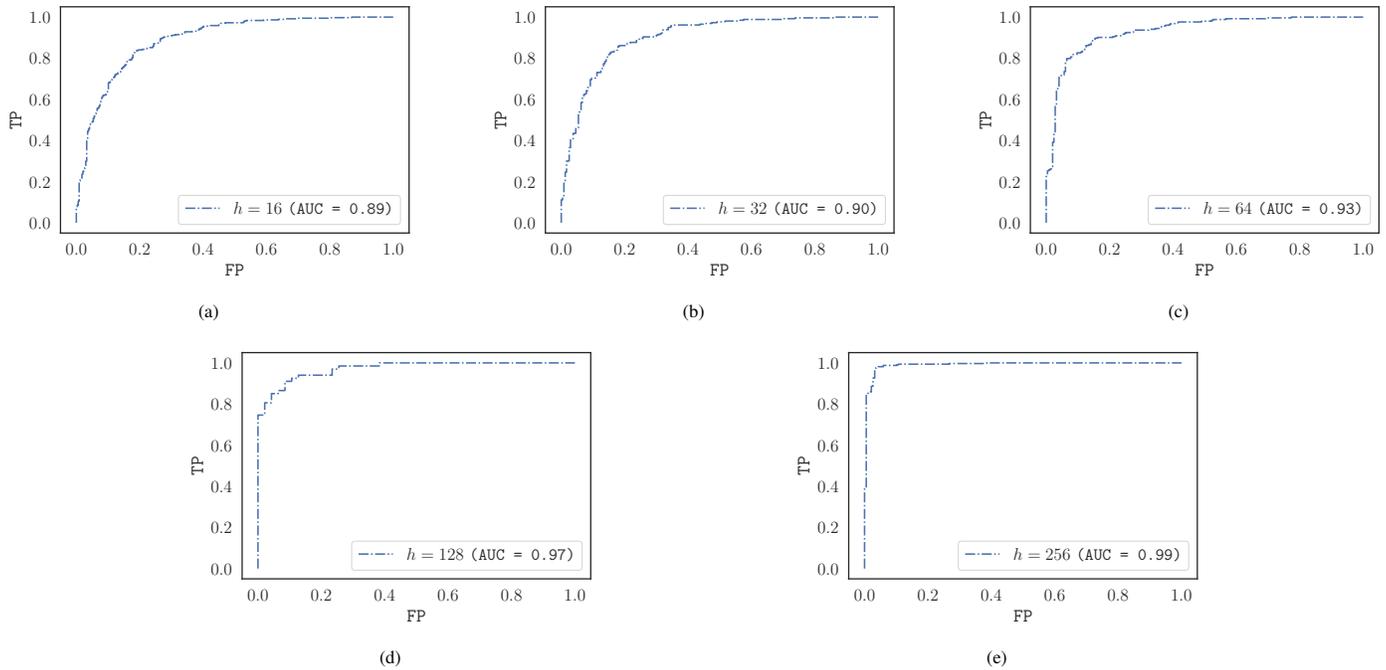


Fig. 4. ROC Curves of the Proposed Model with different Hidden-Layer Size Setting as Follows: (a) $h = 16$, (b) $h = 32$, (c) $h = 64$, (d) $h = 128$, and (e) $h = 256$,

Equation 12), and iii) Recall ($Rec =$ Equation 13), Area Under Curve (AUC), and Receiver Operator Characteristics (ROC) curve.

$$F1 = \frac{TP}{TP + \frac{1}{2}(FP + FN)} \quad (11)$$

$$Pre = \frac{TP}{TP + FP} \quad (12)$$

$$Rec = \frac{TP}{TP + FN} \quad (13)$$

where TP is true positive samples, TN is true negative samples, FP is false positive samples, and FN is false negative samples.

Moreover, Mean Squared Error (MSE) is used to calculate the error rate between predicted values and the actual values. Considering n anomalous sequences inside the dataset, MSE computes the mean of the sum of all the squared errors of each sequence individually (Equation 14).

$$MSE = \sum_{i=1}^n \frac{(a_i - p_i)^2}{n} \quad (14)$$

where a_i is the actual value, and p_i is the predicted value.

3) *Experiment Settings*: The experiments have been conducted using TensorFlow¹. A number of LSTM parameters, including the epochs, batch-size and hidden-layers, have been tested while learning relations and extracting features. Further detail concerning the choice criterion is given in the following section. Recall that batch-size refers to

the size of the embedding vector \vec{v}_i of an activity stream as proposed in the model. The model is trained using Adam optimizer with 0.01 learning rate.

B. Results

The model has been evaluated under different LSTM parameter setting to determine the best performance of relation learning and feature extraction. To this end, we determine the performance of the model under different parameter-setting including epochs, batch-size, and hidden-layer size. To determine the optimal setting for batch-size we conduct a grid search over $\{8, 16, 32, 64, 128\}$. For hidden-layer we also conduct a grid search to tune the best performance over hyper-values of $\{8, 16, 32, 64, 128\}$. The model is run over epochs = 50 and epochs = 100, for the entire grid searches, respectively. Fig. 3 illustrates the performance of the proposed model in terms of MSE value for different parameter setting. The figure shows that epochs = 100 has generally produced better performance than epochs = 50 with different scales of batch-size and hidden-layer. The best results of MSE are recorded with batch-size = 256. It can be seen that the batch-size has an influence on obtaining better results whenever the value get larger although we found that the time complexity is increased accordingly. However, the efficiency on terms of time complexity scale is beyond the scope of this study despite it is interesting for consideration in other simulations such as the case of online feature extraction. Moreover, the model has yield best results whenever hidden-layer = 128 get larger. Figure 4 shows the ROC curves of the proposed model with respect to different hidden-layer sizes. The best result is recorded with AUC = 0.99 at hidden-layer = $h = 256$.

¹<https://www.tensorflow.org/>

TABLE I. THE RESULTS OF ANOMALY BEHAVIOR DETECTION IN TERMS OF F1, PRE AND REC OF THE PROPOSED MODEL COMPARED WITH BASELINES

Method	F1	Pre	Rec
SVM	0.28	34.20	21.20
HMM	0.35	55.98	49.20
NN	0.74	89.76	58.10
Rel-RNN	0.80	99.12	67.12

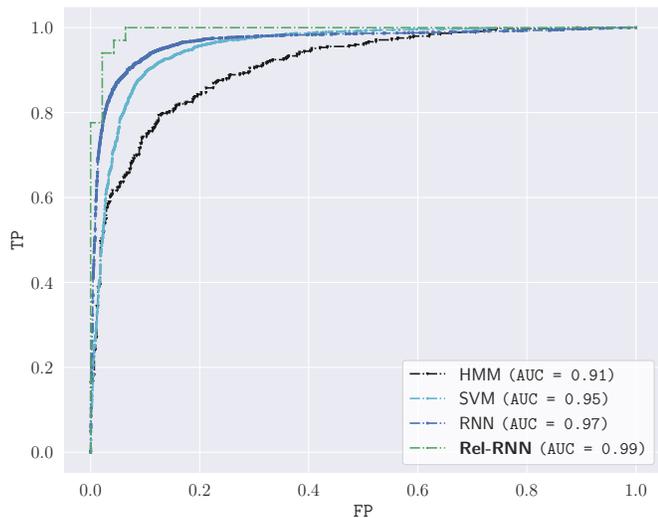


Fig. 5. The ROC Curves that Show the Performance of the Proposed Model along with Baseline Methods.

The proposed model has been evaluated to determine the detection accuracy in terms of F1, Pre and Rec compared with baseline methods. The evaluation explores the detection of anomaly behavior with relation-based feature learning of temporal streams, as in the proposed model, and the aggregated features based on statistician abstraction of audit data. We adopt the following baselines: SVM, HMM, and shallow Neural Network (NN). Note that the later considers feedforward structure with one layer of `batch-size = 265` and `hidden-layer = 128`. Table I shows the obtained results of all methods; for clarity, we denote the proposed method as **Rel-RNN**. It can be seen from the table that the **Rel-RNN** has recorded best results with 0.80, 99.12, 67.12 for F1, Pre, and Rec respectively. To further demonstrate the obtained results, Fig. ?? shows the ROC curve for the proposed model compared with baseline methods. It can be observed that the performance of **Rel-RNN** has obtained a better AUC value of 0.99.

V. CONCLUSION

This study has proposed a novel model for insider threats detection. The proposed model structures the audit data, which represents the daily activities, as a multivariate time series covering broader characteristics for better user behavior learning. Thus, the temporal sequence of exclusive events is considered rather than an abstract set of features. The represented sequences are fed into an RNN model to learn hidden relations for feature extraction. The relations between representative features can be learned to identify latent patterns in the

sequences for recognizing malicious behavior. To maintain the consecutive temporal lags between the set of features, LSTM has been used thus to avoid the vanishing gradient problem. The evaluation on the CERT dataset has shown that the proposed model has outperformed the comparator baselines for insider threat prediction. In the future, the plan is to incorporate Spatio-temporal dependencies to determine whether it affects modeling the latent relations to profile the user's behavior. This is desirable when users have the authorization to access the network from different places remotely. This case is observed during the COVID-19 pandemic when most companies and organizations allow employees to access networks from distant locations.

ACKNOWLEDGMENT

The author would like to thank Al Baha University, Saudi Arabia, for supporting this work under the fund 8/1440.

REFERENCES

- [1] A. B. Pandey, A. Tripathi, and P. C. Vashist, "A survey of cyber security trends, emerging technologies and threats," in *Cyber Security in Intelligent Computing and Communications*. Springer, 2022, pp. 19–33.
- [2] J. Jang-Jaccard and S. Nepal, "A survey of emerging threats in cyber-security," *Journal of Computer and System Sciences*, vol. 80, no. 5, pp. 973–993, 2014.
- [3] Ö. A. Aslan and R. Samet, "A comprehensive review on malware detection approaches," *IEEE Access*, vol. 8, pp. 6249–6271, 2020.
- [4] F. Yuan, Y. Cao, Y. Shang, Y. Liu, J. Tan, and B. Fang, "Insider threat detection with deep neural network," in *International Conference on Computational Science*. Springer, 2018, pp. 43–54.
- [5] S. Yuan and X. Wu, "Deep learning for insider threat detection: Review, challenges and opportunities," *Computers & Security*, p. 102221, 2021.
- [6] M. N. Al-Mhiqani, R. Ahmad, Z. Zainal Abidin, W. Yassin, A. Hassan, K. H. Abdulkareem, N. S. Ali, and Z. Yunos, "A review of insider threat detection: Classification, machine learning techniques, datasets, open challenges, and recommendations," *Applied Sciences*, vol. 10, no. 15, p. 5208, 2020.
- [7] A. Kim, J. Oh, J. Ryu, J. Lee, K. Kwon, and K. Lee, "Sok: A systematic review of insider threat detection." *J. Wirel. Mob. Networks Ubiquitous Comput. Dependable Appl.*, vol. 10, no. 4, pp. 46–67, 2019.
- [8] D. C. Le and A. N. Zincir-Heywood, "Evaluating insider threat detection workflow using supervised and unsupervised learning," in *2018 IEEE Security and Privacy Workshops (SPW)*. IEEE, 2018, pp. 270–275.
- [9] T. Rashid, I. Agrafiotis, and J. R. Nurse, "A new take on detecting insider threats: exploring the use of hidden markov models," in *Proceedings of the 8th ACM CCS international workshop on managing insider security threats*, 2016, pp. 47–56.
- [10] B. Böse, B. Avasarala, S. Tirthapura, Y.-Y. Chung, and D. Steiner, "Detecting insider threats using radish: A system for real-time anomaly detection in heterogeneous data streams," *IEEE Systems Journal*, vol. 11, no. 2, pp. 471–482, 2017.
- [11] D. C. Le and N. Zincir-Heywood, "Anomaly detection for insider threats using unsupervised ensembles," *IEEE Transactions on Network and Service Management*, vol. 18, no. 2, pp. 1152–1164, 2021.
- [12] E. R. DeLancey, J. F. Simms, M. Mahdianpari, B. Brisco, C. Mahoney, and J. Kariyeva, "Comparing deep learning and shallow learning for large-scale wetland classification in alberta, canada," *Remote Sensing*, vol. 12, no. 1, p. 2, 2020.
- [13] A. Tuor, S. Kaplan, B. Hutchinson, N. Nichols, and S. Robinson, "Deep learning for unsupervised insider threat detection in structured cybersecurity data streams," in *Workshops at the Thirty-First AAAI Conference on Artificial Intelligence*, 2017.
- [14] J. Lu and R. K. Wong, "Insider threat detection with long short-term memory," in *Proceedings of the Australasian Computer Science Week Multiconference*, 2019, pp. 1–10.

- [15] S. Paul and S. Mishra, "Lac: Lstm autoencoder with community for insider threat detection," in *2020 the 4th International Conference on Big Data Research (ICBDR'20)*, 2020, pp. 71–77.
- [16] D. S. Berman, A. L. Buczak, J. S. Chavis, and C. L. Corbett, "A survey of deep learning methods for cyber security," *Information*, vol. 10, no. 4, p. 122, 2019.
- [17] A. L'heureux, K. Grolinger, H. F. Elyamany, and M. A. Capretz, "Machine learning with big data: Challenges and approaches," *Ieee Access*, vol. 5, pp. 7776–7797, 2017.
- [18] A. McCarthy, E. Ghadafi, P. Andriotis, and P. Legg, "Functionality-preserving adversarial machine learning for robust classification in cybersecurity and intrusion detection domains: A survey," *Journal of Cybersecurity and Privacy*, vol. 2, no. 1, pp. 154–190, 2022.
- [19] E. T. Ogidan, K. Dimililer, and Y. Kirsal-Ever, "Machine learning for cyber security frameworks: a review," *Drones in Smart-Cities*, pp. 27–36, 2020.
- [20] R. Geetha and T. Thilagam, "A review on the effectiveness of machine learning and deep learning algorithms for cyber security," *Archives of Computational Methods in Engineering*, vol. 28, no. 4, pp. 2861–2879, 2021.
- [21] A. Bendovschi, "Cyber-attacks—trends, patterns and security countermeasures," *Procedia Economics and Finance*, vol. 28, pp. 24–31, 2015.
- [22] A. Sanzgiri and D. Dasgupta, "Classification of insider threat detection techniques," in *Proceedings of the 11th annual cyber and information security research conference*, 2016, pp. 1–4.
- [23] D. C. Le and A. N. Zincir-Heywood, "Machine learning based insider threat modelling and detection," in *2019 IFIP/IEEE Symposium on Integrated Network and Service Management (IM)*. IEEE, 2019, pp. 1–6.
- [24] Y. Wei, K.-P. Chow, and S.-M. Yiu, "Insider threat detection using multi-autoencoder filtering and unsupervised learning," in *IFIP International Conference on Digital Forensics*. Springer, 2020, pp. 273–290.
- [25] M. N. Al-Mhiqani, R. Ahmad, Z. Z. Abidin, W. Yassin, A. Hassan, and A. N. Mohammad, "New insider threat detection method based on recurrent neural networks," *Indones. J. Electr. Eng. Comput. Sci*, vol. 17, no. 3, pp. 1474–1479, 2020.
- [26] T. Hu, W. Niu, X. Zhang, X. Liu, J. Lu, and Y. Liu, "An insider threat detection approach based on mouse dynamics and deep learning," *Security and Communication Networks*, vol. 2019, 2019.
- [27] J. Jiang, J. Chen, T. Gu, K.-K. R. Choo, C. Liu, M. Yu, W. Huang, and P. Mohapatra, "Anomaly detection with graph convolutional networks for insider threat and fraud detection," in *MILCOM 2019-2019 IEEE Military Communications Conference (MILCOM)*. IEEE, 2019, pp. 109–114.
- [28] R. K. Pathan, M. Biswas, and M. U. Khandaker, "Time series prediction of covid-19 by mutation rate analysis using recurrent neural network-based lstm model," *Chaos, Solitons & Fractals*, vol. 138, p. 110018, 2020.
- [29] S. Jeon and J. Moon, "Malware-detection method with a convolutional recurrent neural network using opcode sequences," *Information Sciences*, vol. 535, pp. 1–15, 2020.
- [30] K. Kayama, M. Kanno, N. Chisaki, M. Tanaka, R. Yao, K. Hanazono, G. A. Camer, and D. Endoh, "Prediction of pcr amplification from primer and template sequences using recurrent neural network," *Scientific reports*, vol. 11, no. 1, pp. 1–24, 2021.
- [31] A. F. Agarap, "Deep learning using rectified linear units (relu)," *arXiv preprint arXiv:1803.08375*, 2018.
- [32] C. Wang, K. Viswanathan, L. Choudur, V. Talwar, W. Satterfield, and K. Schwan, "Statistical techniques for online anomaly detection in data centers," in *12th IFIP/IEEE international symposium on integrated network management (IM 2011) and workshops*. IEEE, 2011, pp. 385–392.
- [33] J. Glasser and B. Lindauer, "Bridging the gap: A pragmatic approach to generating insider threat data," in *2013 IEEE Security and Privacy Workshops*. IEEE, 2013, pp. 98–104.

An Improved Label Initialization based Label Propagation Method for Detecting Graph Clusters in Complex Networks

Jyothimon Chandran, V Madhu Viswanatham
School of Computer Science and Engineering
Vellore Institute of Technology
Vellore, Tamilnadu, India

Abstract—Community structure is one of the fundamental characteristics of complex networks. Detection of community structure can provide insight into the structural and functional organization that helps to understand various dynamical processes such as epidemics and information spreading. Label propagation algorithm (LPA) is a well-known method for community structure identification due to linear time complexity. However, the communities extracted by the LPA is unstable since it produces different combinations of communities at each run on the same network. In this paper, a novel label initialization method for label propagation algorithm (ILI-LPA) is proposed to detect stable and accurate community structures. The proposed ILI-LPA focuses on more accurate label initialization rather than assigning unique labels thereby reduce the effect of randomness in LPA. The experiments on several real-world and synthetic networks show that the ILI-LPA improves the quality and stability of communities compared to existing algorithms. The results also demonstrate that appropriate label initialization can significantly improve the performance of label propagation algorithms, and the stability has been improved up to 50-78% relative to the standard LPA.

Keywords—Social networks; community detection; graph clustering; edge clustering coefficient; label initialization; triangle count

I. INTRODUCTION

Complex systems can be modeled as networks, with nodes representing entities of the system and links between nodes denoting its relationships [1]. Such networks are usually termed complex networks and can explain the emergence of complex behavior of the system. Examples of such complex networks [2] are biological networks, citation networks, scientific collaboration networks, and social networks. A common and significant characteristic of complex networks is community structure or communities or clusters [3], such as bacterial communities in the microbial ecosystem and community mobility in urban transport systems. Community structure can provide an overview of the system in consideration, explain the underlying dynamics, and reveal the hidden relations among the entities. It is defined as groups of nodes in a network with dense internal connections inside the groups and fewer connections between the groups [4]. An interesting fact about communities is that the nodes that belong to a community exhibit similar characteristics or common properties that define the overall behavior of the network [5]. Detecting community structure has become an integral part of network analysis.

Several community detection algorithms exist in the literature, and they fall into optimization methods or heuristic methods. The optimization methods such as modularity maximization algorithms [6], [7], [8], spectral methods [9], [10], and evolutionary algorithms [11], [12] formulate an objective function and then estimate an optimal value to find community partitions. Modularity maximization-based methods [6], [7], [8] focus on locating the maximum modularity to extract communities. The spectral methods [9], [10] construct the Laplacian matrix of the network from its characteristic vectors by formulating a quadratic objective function to obtain communities. The methods such as whale optimization [11] and genetic algorithm [12] are evolutionary algorithms. They utilize evolutionary computations to evaluate the optimal value of the optimization function to find the communities. However, most of these algorithms are inappropriate for networks of very large in size because of high time complexity. Heuristic methods apply heuristic techniques to identify communities, which are more time efficient than optimization methods. The methods such as Infomap [13], Edge betweenness [3], [14], and Label propagation algorithm [15] are examples of heuristic community detection algorithms.

The label propagation algorithm (LPA) [15] is one of the computationally efficient community identification methods having time complexity linear ($O(n)$). The LPA consists of mainly two steps: label initialization, and label propagation. At the label initialization, unique labels are assigned to every node in the network. Then at the label propagation, node labels of every node are iteratively updated to the label of the maximum of its adjacent nodes. If more than one label satisfies the maximum criteria, then a random selection on the maximal label is considered. This iterative label update process continues until every node label is the same as its adjacent nodes maximum label. According to a random node order, nodes are processed in each iteration. Finally, the communities are extracted with respect to node labels. Due to the low time complexity, the LPA is a better choice for very large networks. However, the communities detected by the LPA on a network differ in each execution, which makes the algorithm unstable. This drawback prevents LPA from being widely used in practice.

A sample network with two communities with few connections between them is shown in Fig. 1. Nodes {1, 2, 3, 4, 5, 6} and the nodes {7, 8, 9, 10, 11, 12} constitute the first and second communities. Two communities are also

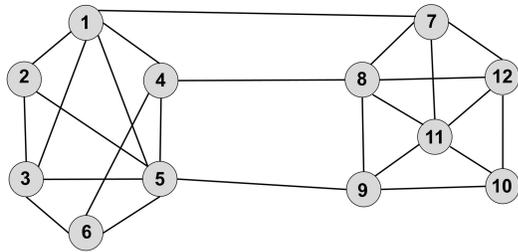


Fig. 1. A Sample Network with Two Communities, $C_1 = \{1, 2, 3, 4, 5, 6\}$, and $C_2 = \{7, 8, 9, 10, 11, 12\}$.

connected through edges (1,7), (4,8), and (5,9), and nodes $\{1, 4, 5, 7, 8, 9\}$ are called boundary nodes. Initially, node numbers are considered as node labels. While initiating the label update process, unique label initialization can cause multiple maximum labels for every node. For instance, if node 4 modifies its label with node 8, and node 11 subsequently changes with the label of node 8, and node 1 updates with node 4, then the algorithm returns a single community. In the next run, if the boundary nodes update their label with the label of nodes in their own community, then it produces two different communities. This is the instability problem of LPA.

To improve the stability of LPA, recently, many improvements have been proposed incorporating measures such as network modularity [16], [17], [18], [19], [20], node strength [21], [22], [23], [24], [25], [26], edge strength [27], [28], [29], [30], [31], [32] and other methods such as node attributes, memory constraints, and evolutionary approaches [33], [34]. Most of these LPA improvements assume that the leading cause of instability is the randomness involved in the label update process and the order of node. Hence eliminating the randomness by incorporating an order to the node selection and label update through various measures were received much attention. The modularity-based label propagation algorithms [16], [17], [18], [19], [20] update node labels according to the label of the neighbor node that produces largest modularity when multiple maximal labels exist. Similarly, node strength-based methods [21], [22], [23], [24], [25], [26] perform label selection according to the node strength measures such as node centrality, influence, or importance during the label update. Similarly, edge strength-based algorithms [27], [28], [29], [30], [31], [32] calculate the strength of connections using measures such as edge clustering coefficient, link strength to update the node label. Overall, these methods focus primarily on providing order to label update rule using various measures, thereby improving the stability of LPA. Though these researches have improved the performance of LPA, there still exists improvements in accuracy and stability.

This paper proposes an improved label propagation algorithm called Identical Label Initialization based LPA (ILI-LPA) based on a novel label initialization method for identifying community structure in networks. Instead of eliminating randomness, ILI-LPA focuses on proper label initialization to improve stability and accuracy. The label initialization of ILI-LPA is based on the measure *triangular structural influence (tsi)*, which estimates influence between nodes based on the triangles in the network. The *tsi* helps to find structurally closely connected nodes in the network to assign identical

labels. Once the label initialization is over, the ILI-LPA follows the random label update strategy to extract the communities.

This paper contains the following contributions.

- Introduces *triangular structural influence (tsi)* to estimate the influence between nodes.
- Proposes a novel label initialization method to tackle stability.
- The effectiveness of ILI-LPA is tested on several real-world and synthetic networks.
- The effect of label initialization on reducing the impact of randomness is assessed.

The rest of the paper is organised as follows: Section II outlines the most recent enhancements to the label propagation algorithm that have been made. Section II elaborates the proposed method. Section IV discusses the experimental details such as data sets, baseline algorithms, and evaluation measures. The results and discussion are presented in section V. Section VI provides the conclusion and future works.

II. RELATED WORK

A complex network is represented in this study by an unweighted undirected network $G(V, E)$, with V denoting the node-set and E denoting the edge set. The neighbor set of node u represents $\Gamma(u)$. If an edge connects two nodes, then they are called neighbors. The degree of node u is represented as d_u . If a node u , ($u \in V$), contains a label, then it is denoted as l_u .

Several LPA improvements were proposed to enhance stability. According to the label update strategy, those LPA improvements can be classified into four. They are modularity-based LPA methods [16], [17], [18], [19], [20], node strength-based methods [21], [22], [23], [24], [25], [26], edge strength-based methods [27], [28], [29], [30], [31], [32], and other LPA improvements [33], [34]. These methods focus mainly on eliminating the randomness to improve the stability and accuracy of the communities produced.

A. Modularity-Based LPA Methods

To improve the stability, Barber and Clark [16] developed LPAm treating the LPA as a modularity optimization problem. According to LPAm, the label to be propagated is the label that increases the modularity. Compared to the original LPA, it improves the quality of the detected communities. This approach, however, has the problem of getting trapped in the local optimum, resulting in incorrect partitions. To avoid local maxima, Liu et al. [17] combined many community pairs at once utilizing a multistep greedy agglomerative algorithm and proposed LPAm+. Both LPAm and LPAm+ have a resolution limit problem due to the modularity function, which also increases the time complexity. To address the shortcomings of LPAm+, Le et al. [19] presented an improved LPAm+ algorithm called meta-heuristic-based LPA, which was based on the Record-to-Record Travel algorithm. This algorithm improves modularity prior to community merging. Another improved LPA called Stepping LPA-S was proposed by Li et al. [18] in which labels are propagated based on similarity. The stepping

LPA-S picks the label that results in the highest modularity. A modularity gain acceleration method based on modularity was introduced in [20] by formulating an objective function. The objective function is solved using global and local sum weights. Each node's label transition is computed using local sum and general sum is calculated for each label. However, because of the modularity function, the time complexity of the above-mentioned algorithms is significantly higher than the other LPA improvements. Therefore, these are unsuitable for very large-scale networks.

B. Node Strength-Based LPA Methods

The idea of node strength-based label propagation algorithms is that when multiple labels satisfy the maximum criteria, instead of a random selection, the label of the node with the highest importance is chosen to overcome the instability problem. More crucially, calculating each node's importance in the network is the main task of these methods. Xing et al. [21] put forward the NIBLPA method utilizing the k-shell value to determine which label to update. The influence of nodes is assessed by examining the nodes degree and k-shell value along with its neighbours k-shell values. Subsequently, Zhang et al. [22] proposed LPA_NI, which considers both node importance and label influence. LPA_NI first estimates node importance using both the node's priori influence and the degree and the influence of its neighbors. The algorithm computes the influence of each label and updates the node label with the most influential label. Tasgin and Bingol [23] presented a local approach based on label propagation for detecting communities via boundary node identification. This approach first finds and rank the boundary nodes. Subsequently, the label of the node that has the largest score among its neighbours is spread. A method (NI-LPA) based on node importance was suggested in [24]. The node importance was calculated considering each node's signal propagation capability, Jaccard distance and k-shell value. However, the time complexity is increased to $O(n^2)$. The paper [25] employed label importance and proposed a label importance-based LPA (LILPA). The label update process in LILPA depends on the importance and attraction of nodes and label importance. The LILPA follows a fixed node order in which the nodes are arranged according to node importance, calculated from using closeness and degree of nodes. Incorporating the modularity and node significance, Li et al. [26] presented an enhanced algorithm called LPA_MNI. It begins by initializing each node with a unique community. Following that, a rough community is built for every node based on modularity gain by merging each node with its neighbor community in descending order of node importance until no further improvement is possible. The node strength is quantified using normalized degree centrality.

C. Edge Strength-Based Label Propagation Methods

Some researchers considered the strength of connections between nodes (edges/links) rather than the node importance to identifying community structure in networks. Based on edge strength, Lou et al. [27] introduced LPA_CNP algorithm. To begin, this method calculates the weighted coherent neighborhood propinquity for each pair of nodes to reflect the chance that two vertices are members of the same community. A node's label is updated to the label with the

highest weighted-CNP. The results indicate that LPA_CNP outperforms LPA, particularly in large-scale networks. Zhang et al. [28] suggested an edge clustering-based LPAc algorithm. It first calculates the edge clustering coefficients of every edge in the network. During label propagation, this strategy selects the label of largest edge clustering coefficient edge that connects the neighbor. According to the link influence and node strength, Berahmand and Bouyer [29] presented LP_LPA. This approach initially determines the similarity of links between nodes assuming that nodes within a community share more common neighbors than nodes in other communities. Therefore, node strength is also estimated according to degree centrality, and initial node selection is performed by calculating node strength. Jokar and Mosleh [30] proposed BLDLP, which determines the weight for each edge according to the link density. If nodes' maximum labels are not unique, the largest weight edge label is chosen. Jiang et al. [31] introduced a link similarity measure and proposed LLPA in which node labels were computed from the link weights to identify functional modules. Li et al. [32] studied the network's higher-order properties by determining the most representative triangle motif that encoded the strength of connections and presented a community recognition approach based on Motif-Aware Weighted Label propagation. As a result, a unique voting approach termed NaS is presented to reduce the randomness provided by tie-breaking.

D. Other Label Propagation Methods

Hosseini and Rezvani [33] introduced the AntLP based on ant colony optimization (ACO). The algorithm begins by weighting all the edges employing a combination of similarity indexes. Then, it attempts to spread labels by grouping comparable vertices to optimize modularity of each community according to their similarity of vertices. Berahmand et al. [34] proposed SAS-LP algorithm, an improved LPA algorithm for attributed graphs that addresses issues of instability and low quality while maintaining structural cohesiveness and attribute homogeneity in the detected communities.

Many of the LPA improvements, discussed in Sections 2.3 and 2.4, recommend strategies or techniques on label update procedure and node order, and follow a standard unique label initialization to improve the stability. The significance of label initialization on improving the stability and accuracy still remains an open problem. The aim of the paper is also on evaluating the impact of identical label initialization instead of unique label initialization on the stability and accuracy of LPA.

III. PROPOSED METHOD: ILI-LPA

This section proposes a novel label initialization method for LPA to extract community structure in networks. The proposed algorithm is named Identical Label Initialization based Label Propagation Algorithm (ILI-LPA). Different from the standard LPA and its improvements described in section 2, the ILI-LPA focuses on appropriate label initialization to identify stable and accurate communities.

The effect of an appropriate label initialization is illustrated in Fig. 2. Assume that some of the nodes in each community is assigned with same labels. The label of nodes in each set

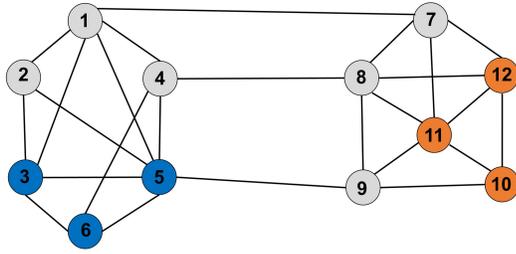


Fig. 2. A Simple Network with Two Communities in which the Nodes {3, 5, 6} Share a Common Label, Nodes {10, 11, 12} also Share another Label and the Remaining Nodes have different Labels.

{3, 5, 6} and {10, 11, 12} is same and the remaining nodes ({1, 2, 4, 7, 8, 9}) carries unique labels. Since some of the nodes in each community is assigned with the same labels, the stability of the label propagation improves significantly. This is because the boundary nodes ({1, 4, 5} and {7, 8, 9}) can never update their label to the node labels that lie in other community based on the label propagation rule. One can see that nodes {1, 4, 5} can never update their labels to the label of other community nodes because their maximum neighbors' labels lie within the community. This applies to the nodes {7, 8, 9} also. when the random label update is applied. The main idea of this paper is to find such nodes that possess a high probability of joining a single community and assigning the same labels to them, thereby improving stability and accuracy.

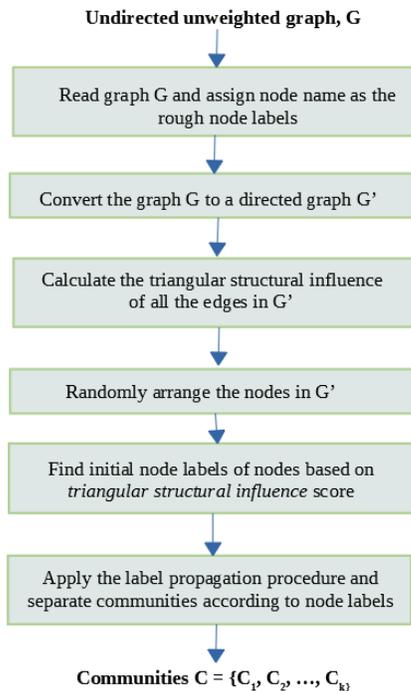


Fig. 3. The Block Diagram the of the Proposed ILI-LPA Method.

The proposed algorithm consists of mainly two phases: identical label initialization and label propagation. At the label initialization phase, ILI-LPA finds structurally closely connected nodes and assigns identical labels to them. When a node is connected with most of the neighbors of a neighbor

node, then the node is said to be structurally closely connected to the neighbor node. To find the nodes that are structurally closely connected, a local measure called *triangular structural influence (tsi)* is introduced from the idea of edge clustering coefficient [36]. Then, according to the *tsi*, the label of the most influential node is given to its structurally closely connected neighbors to initialize node labels. Once the label initialization is over, the ILI-LPA performs the label propagation in which node labels are updated to the neighbors' maximal label. If there exist multiple maximum labels, then a random label selection strategy has opted. Each steps of the proposed ILI-LPA is provided in Fig. 3.

A. Identical Label Initialization

At the label initialization phase, the ILI-LPA aims to assign the same labels to structurally closely connected nodes. It can be measured by estimating the strength of connections between nodes. Several similarity measures exist in the literature that quantifies the connection strength between nodes. These measures quantify similarity considering the network structure or topology to reveal the strength of connections. It includes:

Definition 1: Cosine similarity [35] defined for node i and j is:

$$CS(i, j) = \frac{|\Gamma(i) \cap \Gamma(j)|}{\sqrt{d_i \cdot d_j}} \quad (1)$$

Definition 2: Jaccard Similarity [35] defined for node i and j is:

$$JS(i, j) = \frac{|\Gamma(i) \cap \Gamma(j)|}{|\Gamma(i) \cup \Gamma(j)|} \quad (2)$$

Definition 3: Sørensen Index [35] defined for node i and j is:

$$SS(i, j) = \frac{2 \cdot |\Gamma(i) \cap \Gamma(j)|}{d_i + d_j} \quad (3)$$

Definition 4: Hub Depressed Index [35] defined for node i and j is:

$$HDI(i, j) = \frac{|\Gamma(i) \cap \Gamma(j)|}{\max\{d_i, d_j\}} \quad (4)$$

Definition 5: Hub Promoted Index (HPI) [35] defined for node i and j is:

$$HPI(i, j) = \frac{|\Gamma(i) \cup \Gamma(j)|}{\min\{d_i, d_j\}} \quad (5)$$

Definition 6: Edge clustering coefficient (ECC) [36] is:

$$ECC(i, j) = \frac{|\Gamma(i) \cup \Gamma(j)| + 1}{\min\{d_i - 1, d_j - 1\}} \quad (6)$$

where d_i and d_j indicate degrees of node i and j , $\Gamma(i)$ signifies the neighbors of node i , $|\Gamma(i) \cap \Gamma(j)|$ estimates the number of common nodes. However, these measures are unidirectional, which means that they assume the influence between nodes are equal. In reality, the strength between nodes is bidirectional. Therefore, a new measure is introduced to measure the influence between nodes.

These similarity measures estimate the (structural) strength of connections between nodes accurately. More importantly, it is also known that high similarity value between nodes indicates same community participation of nodes. However, when the (structural) strength of relationship between nodes are considered, one can see that node strength from node i to j and vice versa may not be always same. If an edge that connects a pair of nodes is densely connected by their neighbors, then the edge clustering coefficient of that edge will be comparatively larger. This shows that the nodes exhibit high probability to lie in the same community. From the idea of the edge clustering coefficient, we introduce *triangular structural influence (tsi)* to quantify the node strengths. The *tsi* estimates the strength of the relationship (influence) between nodes based on the triangles associated with each node.

Definition 8: *triangular structural influence tsi(i, j)* denotes the influence the node i exerts on node j . If (i, j) is an edge in the network, $tsi(i, j)$ is the ratio of the actual number of connections from node i to the neighbors of j to the maximum possible connections. Therefore $tsi(i, j)$ is defined as:

$$tsi(i, j) = \frac{1 + |\Gamma(i) \cap \Gamma(j)|}{|\Gamma(j)|} \quad (7)$$

where $\Gamma(i)$ represents the neighbor set of i , $|\Gamma(j)|$ denotes the number of neighbors of node j (degree of j), $|\Gamma(i) \cap \Gamma(j)|$ indicates the number of common neighbors of node i and j , i.e., it represents the number of triangles that connects node i and j . If a node i is connected to all the neighbors of node j , then there exists high influence from i to j .

Similarly, the *tsi* from node i to node j cannot be same as the *tsi* from node j to node i . Therefore $tsi(j, i)$ is:

$$tsi(j, i) = \frac{1 + |\Gamma(i) \cap \Gamma(j)|}{|\Gamma(i)|} \quad (8)$$

The densely connected nodes express high *tsi* than loosely connected nodes. Thus, it is clear that the nodes in a community exhibit higher *tsi* between nodes and less influence between communities.

Once the *tsi* between nodes are estimated, unique rough labels are assigned to every node in the network. Subsequently, in order to assign initial labels (9) is used. During the identical label initialization, each node's label is updated to the label of its neighbors based on the *tsi* using (9).

$$l_j^{init} = \arg \max_{i \in \Gamma(j)} L(l_i, l_j) \cdot f(i, j) \quad (9)$$

where l_j^{init} denotes the initial label of node j , $f(i, j)$ is a function that returns 1 if $tsi(i, j) \geq tsi(j, i)$ and $tsi(i, j) \geq \Theta$. L denotes the label. The Equation (9) can also be interpreted that every node in the network try to update the label of each of its neighbor node with its own label, if $f(i, j)$ is satisfied.

B. Label Propagation

During the label propagation, each node's label is updated asynchronously at random to the label shared by the majority of its neighbors. The proposed method (ILI-LPA) updates the

labels in the same way as the standard LPA does. The every node label is updated using (10).

$$L_i = \arg \max_l |\Gamma^l(j)| \quad (10)$$

where $\Gamma^l(i)$ denotes the neighbors of node i with label l . Communities are formed through an iterative process in which densely connected groups of nodes reach consensus on a single label. Finally, the method converges when there are no more changes to the nodes' labels. If there exists more than one maximal label, the ILI-LPA follows the same LPA strategy, in which ties are broken randomly. Finally, nodes are categorized according to the node labels. That is, nodes with same labels join the same community. The Algorithm 1 describes the procedure of ILI-LPA in detail.

Algorithm 1 The Proposed ILI-LPA

Input: Undirected network $G = (V, E)$, parameter θ
Output: Communities $C = \{C_1, C_2, C_3, \dots, C_k\}$

```

1: procedure ILI-LPA( $G, \Theta$ )
2:   Read network  $G$   $\triangleright$  Phase 1: Label Initialization
3:   Assign rough unique labels to every node in  $V$ 
4:   Convert the network  $G$  to directed network  $G'$  by
   adding directions
5:   for each edge in  $G'$  do
6:     Calculate tsi of the edge
7:     Attach it as edge weights
8:   end for
9:   Arrange the nodes in  $V$  in random order
10:  for each node  $u$  in  $G'$  do
11:    for each node  $v \in \Gamma(u)$  do
12:      if  $tsi(u, v) \geq tsi(v, u)$  and  $tsi(u, v) \geq \Theta$  then
13:        update the label of  $v$  with the label of  $u$ 
14:      end if
15:    end for
16:  end for
17:  Removes the directions and weights of  $G'$ 
18:  Set  $t=1$   $\triangleright$  Phase 2: Label propagation
19:  Arrange the nodes  $V$  in random order and set it to  $V'$ 
20:  for each node  $u$  in  $V'$  do
21:    update its label according to equation (10)
22:    if there exists more than one maximum label then
23:      randomly update to the label of maximum of
   its neighbors.
24:    end if
25:  end for
26:  goto step 27 if none of the node label changes, else
   set  $t = t+1$ , go to step 19
27:  According to the node label, separate the communities.
28:  Return communities  $C$ .
29: end procedure

```

The algorithm first assigns unique rough labels all the nodes. In step 8, it converts the input (undirected) graph to a directed graph by adding directions to all edges. Then, at step 5-8, the *tsi* of each edge is computed using (7) and (8) and attach it as corresponding edge weights. Subsequently, each node tries to spread its label to each of its neighbors and update the neighbor's label with its label if the *tsi* from the node to its neighbor is greater than the neighbor back

to the node. The label spread and label update is performed sequentially in an asynchronous fashion to the entire nodes in the network at once. It is performed in step 18-24. Before that, the algorithm converts the network to an undirected network and remove the edge weights and retain only the node labels. The network contains nodes with identical labels in densely connected regions. This process is followed by the label propagation to find final communities. At step 23, node labels are updated according to the maximum label of their neighbors.

The main steps of ILI-LPA is illustrated in Fig. 4 with the support of a toy network. Fig. 4 (a) shows a network that contains three communities $\{1, 2, 3, 4, 5, 6\}$, $\{7, 8, 9, 10, 11\}$, $\{12, 13, 14, 15, 16, 17, 18\}$. Figure 3 (b) shows the estimated tsi to between nodes and marked at the ends of each edge which represents the tsi to that node. The initial community labels identified by the proposed method by the label initialization is represented in Fig. 4 (c), where the value associated with each node indicates that the node label. Fig. 4 (d) indicates the communities identified after the label propagation. The nodes and their corresponding community labels are given to express the node and its updated nodes label.

C. Complexity Analysis

The ILI-LPA contains mainly two phases. The major steps in phase 1 are unique label initialization and *triangular structural influence* estimation. It takes $O(n)$ time to initialize rough unique labels to every node in the network. The time complexity of estimating the tsi of every edge in both directions is $O(m.d_{avg})$ where d_{avg} and m denote the average degree and the number of edges. The label propagation takes $O(m)$ time. Thus the overall time complexity of ILI-LPA is $O(m.d_{avg}) + O(n) + O(m)$, which is approximately equal to $O(m.d_{avg}) \approx O(n.d)$ ($d \ll n$).

IV. EXPERIMENTAL DETAILS

The performance of the ILI-LPA was tested on synthetic networks and real networks. Experiments were carried out on a 3.4 GHz Intel Core i7 CPU with 16.0 GB of RAM. Python-Networkx was used to develop the algorithm. The algorithm's input parameter θ is set at 0.35 on all the networks.

A. Datasets

1) *Real-World Networks*: The networks considered in this study are: Karate Club [37], Dolphin network [38], Football [3], Polbooks [39], Netscience [40], Email Enron [41], Cond-mat-2003 [41], Cond-mat-2005 [41], DBLP [42], Amazon [42]. The details of these networks are provided in Table I [N_C : actual communities in the network, d_{avg} : average degree, CC : Clustering coefficient].

B. Synthetic Networks

Lancichinetti-Fortunato-Radicchi (LFR) [43] is a popular synthetic network generator to test the performance of community detection algorithms. The community size and degree distributions of generated networks follow power-law distributions. The LFR generator contains the number of nodes

TABLE I. DETAILS OF THE REAL-WORLD NETWORKS

Dataset	Nodes	Edges	NC	d_{avg}	CC
Karate Club	34	78	2	4.58	0.58
Dolphin	62	159	2	5.13	0.30
Football	115	613	12	10.66	0.40
Polbooks	105	441	3	8.40	0.49
Netscience	1589	2742	-	3.451	0.878
Email-enron	36692	183831	-	10.02	0.716
Cond-mat-2003	31163	120029	-	7.703	0.723
Cond-mat-2005	40421	175692	-	8.693	0.719
Amazon	334863	925872	-	5.530	0.396
DBLP	317080	1049866	-	6.662	0.632

TABLE II. THE PARAMETER VALUES OF THE LFR NETWORK

Network name	N	k	$kmax$	cm_{ax}	t_1	t_2	μ
LFR_net1	1000	20	100	100	2	1	0.05 - 0.75
LFR_net2	5000	20	500	500	2	1	0.05 - 0.75
LFR_net3	10000	20	1000	1000	2	1	0.05 - 0.75
LFR_net4	20000	20	2000	2000	2	1	0.05 - 0.75

(N), maximum degree ($kmax$), maximum community size (cm_{ax}), average degree (k), degree distribution exponent (t_1), community size distribution exponent (t_2). The most important parameter that sets the character of communities is the mixing parameter μ , which indicates the percentage of linkages between communities and within communities. Table II shows the parameter values given to generate LFR network.

C. Baseline Algorithms

The proposed algorithm was compared with seven existing algorithms. They are Fastgreedy [6], Louvain [7], Infomap [13], LPA [15], NIBLPA [21], LPA-CNP-E [27] and Stepping-LPA [18].

D. Evaluation Metrics

Modularity: For assessing the quality of community partitions, the modularity [44] metric is widely used. Modularity (Q) indicates the percentage of edges within the communities minus the expected percentage of community edges of a random network with same degree distribution. The modularity value of communities C is computed using (11).

$$Q(C) = \frac{1}{2m} \sum (A_{ij} - \frac{k_i k_j}{2m}) \delta(C_i, C_j) \quad (11)$$

where m signifies the total edges, A denotes the adjacency matrix representation of the network, $A_{ij} = 1$ if node i and j are connected, 0 otherwise. The k_i and k_j indicate the degrees of node i and j . The C_i and C_j denote the communities of nodes i and j . The Kronecker delta (δ) yields 1 when nodes i and j belong to a single community, otherwise, it returns 0.

Normalized Mutual Information (NMI): The NMI [45] measures how similar the communities are to one other, by comparing the communities extracted by the community detection algorithm to the actual communities. Let A and B be the actual and detected communities of a given network. The NMI (A, B) is computed using (12).

$$NMI(A, B) = \frac{-2 \sum_{i=1}^{C_A} \sum_{j=1}^{C_B} N_{ij} \log \frac{N_{ij} N}{N_i N_j}}{\sum_{i=1}^{C_A} N_i \log \frac{N_i}{N}} \quad (12)$$

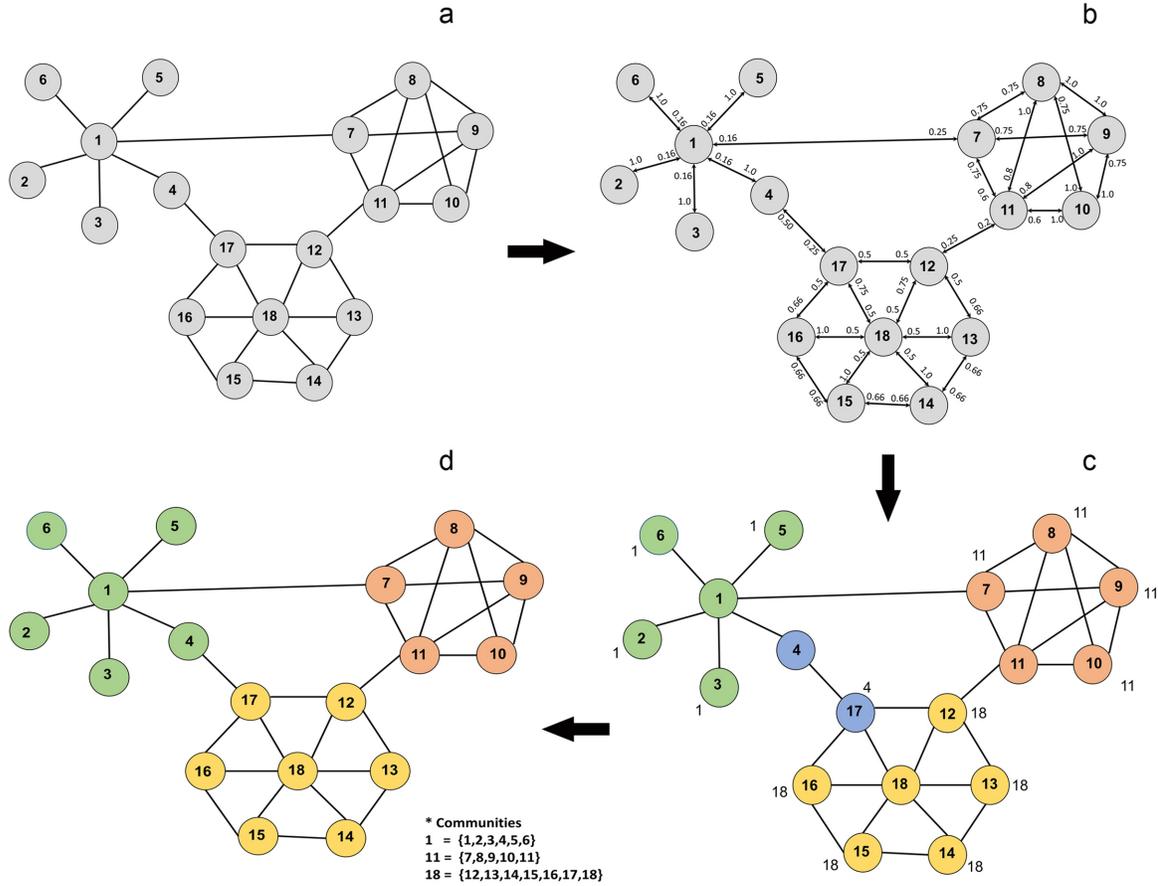


Fig. 4. Details of each Major Steps of ILI-LPA with an Example. a) Input Network, b) Triangular Structural Influence as Weights to each Edge, c) Initial Labels of the Nodes, d) the Final Extracted Communities.

where N_{ij} is the number of common nodes of A's community i and B's community j . The actual and discovered number of communities are denoted as C_A and C_B , respectively. NMI produces its maximum value of 1 if the discovered partition is identical to the actual partitions. If the two partitions are not related, the NMI returns 0.

V. EXPERIMENTAL RESULT AND DISCUSSION

A. Comparing the Modularity of Algorithms on Small Networks

The modularity of the ILI-LPA and the compared algorithms on small networks is presented in Table III. All algorithms were run 100 times, and calculated the average and standard deviation. From Table 3, we can see that ILI-LPA produces significantly better modularity in comparison with baseline algorithms, including the Louvain method on Dolphin and Football networks. On Polbooks network, the ILI-LPA algorithm produces modularity value of 0.526, which is closer to the modularity of Louvain. Though the obtained average modularity of ILI-LPA is only 0.371 on Karate dataset, but still, it is better than standard LPA.

TABLE III. MODULARITY OF ALGORITHMS ON SMALL REAL-WORLD NETWORKS

Algorithms	Karate	Dolphin	Football	Polbooks
Fastgreedy	0.381	0.495	0.568	0.502
Louvain	0.415	0.519	0.604	0.527
Infomap	0.415	0.520	0.563	0.512
LPA	0.357	0.487	0.589	0.511
LPA-CNP-E	0.303	0.463	0.601	0.451
Step-LPA-S	0.371	0.378	0.575	0.496
NIBLPA	0.40	0.43	0.50	0.55
ILI-LPA	0.371±0.00	0.523±0.05	0.604±0.02	0.526±0.002

B. Comparing the NMI of Algorithms on Small Networks

To evaluate the accuracy of algorithms, the NMI is calculated and reported in Table IV. It provides that on Football and Polbooks networks, the NMI score of ILI-LPA is comparatively higher compared to baseline algorithms. Though the modularity value of Dolphin network shown in Table 3 is high, the accuracy is low, with NMI value 0.566. On Karate network, ILI-LPA algorithm yields an NMI score of 0.837, which is just below the NMI of Step-LPA-S and better than other methods, including Louvain. From Tables III and IV, we can say that the ILI-LPA is better than the others on stability and accuracy on small networks.

TABLE IV. NORMALIZED MUTUAL INFORMATION OF ALGORITHMS ON REAL-WORLD NETWORKS

Algorithms	Karate	Dolphin	Football	Polbooks
Fastgreedy	0.693	0.573	0.744	0.439
Louvain	0.707	0.474	0.885	0.418
Infomap	0.707	0.563	0.921	0.467
LPA	0.649	0.540	0.893	0.524
LPA-CNP-E	0.837	0.731	0.909	0.571
Step-LPA-S	0.924	0.888	0.925	0.571
NIBLPA	0.58	0.50	0.72	0.53
ILI-LPA	0.837	0.566	0.927	0.593

TABLE V. MODULARITY OF ALGORITHMS ON LARGE REAL-WORLD NETWORKS

Algorithms	Net science	Email-enron	Cond-mat -2003	Cond-mat -2005	Amazon	DBLP
Fastgreedy	0.955	0.510	0.678	0.631	0.879	0.728
Louvain	0.959	0.605	0.761	0.722	0.910	0.810
Infomap	0.931	0.527	0.661	0.631	0.232	0.714
LPA	0.912	0.337	0.592	0.620	0.784	0.634
LPA-CNP-E	0.932	0.512	0.736	0.631	-	-
Step-LPA-S	0.921	0.531	0.694	0.625	-	-
NIBLPA	0.68	0.12	0.50	0.23	0.67	0.61
ILI-LPA	0.921	0.562	0.632	0.645	0.813	0.653

C. Comparing the Modularity of Algorithms on Large Networks

Table V provides the modularity produced by the algorithms on large networks. As seen on the Table, the modularity value of the proposed algorithm on the network science dataset is 0.921, significantly better than LPA, Step-LPA-S, NIBLPA. On Email-enron network, the ILI-LPA gives 0.562 modularity, which is higher than all the algorithms except the Louvain method. On Cond-mat-2003 dataset, our algorithm is not performing well because the modularity of the detected communities is just 0.632. At the same time, On Cond-mat-2005 network, the modularity of the proposed method is superior to other algorithms except for Louvain. On Amazon, only Fastgreedy and Louvain produces superior modularity than the proposed method. On DBLP network, though our algorithm is inferior to non-LPA-based algorithms, still better than both LPA and NIBLPA. The experiments on large-scale networks compared to LPA-based (LPA, LPA-CNP-E, Step-LPA-S, NIBLPA) and non-LPA-based (Fastgreedy, Louvain, Infomap) algorithms on modularity metric show that the ILI-LPA has better performance on LPA-based algorithms and is closer to non-LPA based algorithms except Louvain. Since Louvain is a modularity optimization method, Louvain can return higher modularity on most of the network. Table 5 demonstrates that the ILI-LPA performs well on large-scale real-world networks.

D. Evaluating the Performance ILI-LPA on LFR Networks

Extensive tests have been carried out on the LFR network in order to validate the performance of the ILI-LPA. It is analyzed on the LFR network in three different aspects: modularity, NMI, and the number of communities. Since the actual community information is available in the LFR network, the actual modularity and number of communities of the corresponding network are considered as GroundTruth value. The algorithms employed for the comparison are Fastgreedy, Louvain, Infomap, and LPA. Since non-LPA algorithms are

better than LPA variants, only these four algorithms are considered for synthetic network evaluation. The results with respect to Modularity (Q) (including the GroundTruth modularity of LFR communities) is illustrated in Fig. 5. With an increase in the mixing parameter (μ), the accuracy of algorithms steadily diminishes. Fig. 5 shows that the modularity of the communities identified by the proposed algorithm is the same as that of the GroundTruth modularity of LFR until reaches 0.70. Though the modularity of Fastgreedy is stable, it is significantly lower than the GroundTruth.

Fig. 5 shows that on all the four networks, the ILI-LPA produces modularity closer to the GroundTruth and better than the other LPA methods. When the μ is less than 0.45, the modularity of detected communities of different algorithms, except the Fastgreedy, is closer to GroundTruth modularity on all four networks. Both LPA and Infomap modularity drastically reduce when at 0.50 and 0.55, respectively. However, ILI-LPA is the same as GroundTruth and Louvain communities until the reaches 0.70. when the modularity of standard LPA drops between 0.40 and 0.50, the proposed improved LPA (ILI-LPA) algorithm maintains the quality of communities till reaches 0.70. In all four networks in Fig. 5, the ILI-LPA shows a similar pattern of modularity and better performance than other algorithms.

The experimental result with respect to NMI is reported in Fig. 6. The results indicate that except the FastGreedy, all algorithms produce good performance on all the four networks until is 0.45. With respect to the increase in the mixing parameter (μ), the difficulty in community identification also increases. Fig. 6 shows that the proposed method produces stable and accurate communities on all the four LFR networks until is less than 0.7. The performance of Fastgreedy and Infomap significantly decreases when is above 0.4, and the performance of LPA also drastically drops when crosses 0.5. The experiments on four LFR networks demonstrate that the ILI-LPA is improved in terms of algorithms than compared algorithms.

Additionally, while evaluating the performance of community detection methods, the number of communities detected is a significant performance metric to consider. In LFR networks, since the actual number of communities is known, the comparison can provide more insights into the performance. The number of communities produced by the algorithms corresponding to four different μ values is illustrated in Fig. 7. In LFR net1, the number of communities of the ILI-LPA is very close to the GroundTruth communities of the LFR network in all the four different μ values. On all four networks, the number of communities of Fastgreedy is significantly low than the GroundTruth values. While μ is greater than 0.4, as expected the both Infomap and LPA yield poor performance. There is only one algorithm that produces an exact number of communities to the ground truth in all the test cases is our proposed ILI-LPA algorithm. Though the Louvain shows high modularity value, the identified community size of Louvain is significantly smaller than that of the actual number of communities. Overall, the experimental results illustrate that the ILI-LPA is stable and accurate in finding communities without consuming much computational time.

Analysis and Discussion The results illustrate that the ILI-LPA improves the stability and accuracy without significantly

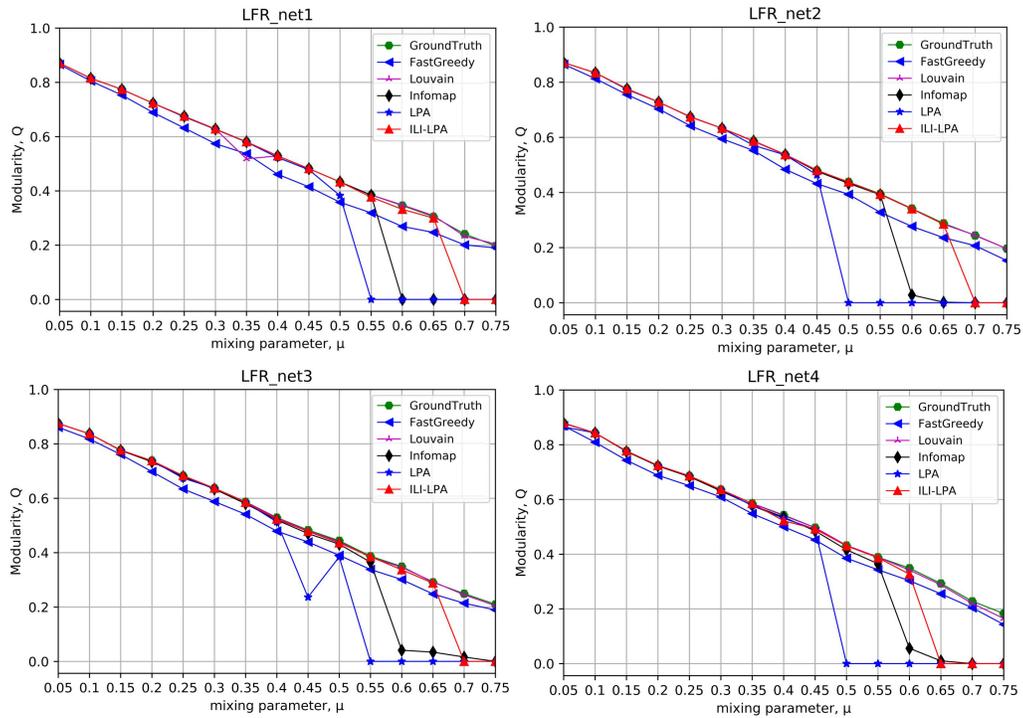


Fig. 5. Modularity of the Communities Detected by the Five Community Detection Algorithms along with the Actual Modularity (GroundTruth) on Four LFR Networks.

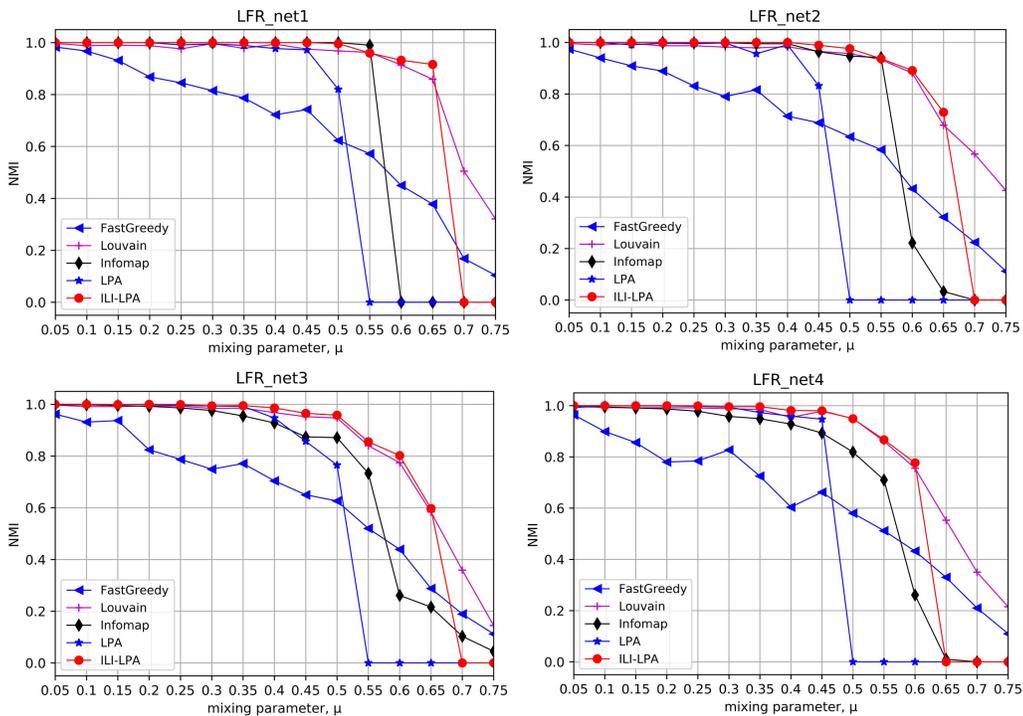


Fig. 6. NMI Produced by the Algorithms on Four LFR Networks.

increasing the execution time. Unlike the standard and other improvements of LPA that assign nodes with unique labels during the label initialization, the ILI-LPA focuses on identical

label initialization where two nodes get the same label if the nodes are structurally closely connected. That is, the two nodes expresses a high probability to continue in a single community

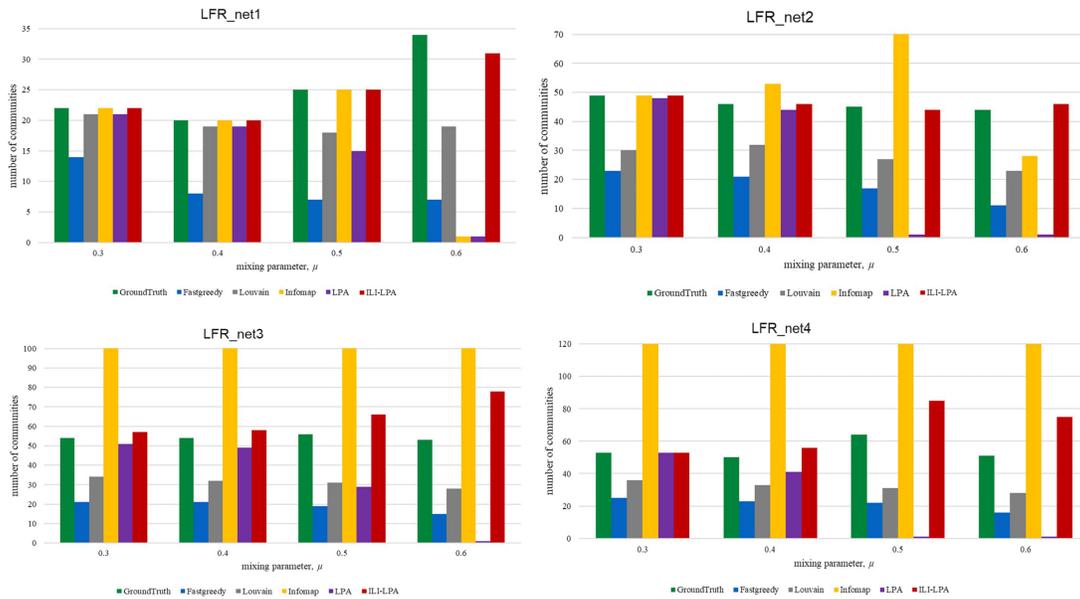


Fig. 7. The Number of Communities Produced by the Algorithms on Four LFR Networks.

from the initialization to the final communities during the label propagation. Real-world and synthetic networks are employed for conducting the experiments. The results prove the importance of more accurate label initialization than the conventional unique label initialization to improve the stability and accuracy of improved LPAs. The main advantage of ILI-LPA is that the identical initialization of labels reduces number of iterations at the label propagation phase. Also, the label initialization helps each node to differentiate its own community neighbors from other neighbors, which solves the instability due to random selection. So that, without eliminating the randomness in LPA and employing proper label initialization, the ILI-LPA achieves better performance.

VI. CONCLUSION AND FUTURE WORK

The LPA is a popular time-efficient community detection algorithm. However, the instability in results is its main drawback. Many of the recent LPA improvements concentrate primarily on eliminating randomness by introducing various measures that provide an order to the label update process. However, these improvements either increase the computational time or reduce the accuracy. This paper presents a novel technique called identical label initialization to improve stability and accuracy and proposes the ILI-LPA. The ILI-LPA finds nodes that are structurally closely connected, then assigns identical labels to them. Our approach focuses primarily on proper label initialization rather than assigning unique labels. The ILI-LPA maintains the random label selection strategy when multiple maximal labels exist to update node labels. The results demonstrate that the proposed ILI-LPA has better performance than the existing algorithms. The results also demonstrate that proper label initialization is also an important factor for improving LPA stability and accuracy. In future, the proposed method can be extended for finding overlapping communities in networks. In addition, the label initialization can be extended to detect evolving communities in dynamic networks.

REFERENCES

- [1] M. E. J. Newman, "The structure and function of complex networks," *SIAM Review*, vol. 45, no. 2, pp. 167–256, 2003.
- [2] M. E. J. Newman, "Networks: An Introduction," *J. Math. Sociol.*, 2013.
- [3] M. Girvan and M. E. J. Newman, "Community structure in social and biological networks," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 99, no. 12, pp. 7821–7826, 2002.
- [4] G. Xu, J. Guo, and P. Yang, "TNS-LPA: An Improved Label Propagation Algorithm for Community Detection Based on Two-Level Neighbourhood Similarity," *IEEE Access*, vol. 9, pp. 23526–23536, 2021.
- [5] S. Kumar, L. Singhla, K. Jindal, K. Grover, and B. S. Panda, "IM-ELPR: Influence maximization in social networks using label propagation based community structure," *Appl. Intell.*, vol. 51, pp. 7647–7665, 2021.
- [6] A. Clauset, M. E. J. Newman, and C. Moore, "Finding community structure in very large networks," *Phys. Rev. E - Stat. Physics, Plasmas, Fluids, Relat. Interdiscip. Top.*, vol. 70, no. 6, p. 6, 2004.
- [7] V. D. Blondel, J. L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," *J. Stat. Mech. Theory Exp.*, vol. 2008, no. 10, 2008.
- [8] Z. Bu, C. Zhang, Z. Xia, and J. Wang, "A fast parallel modularity optimization algorithm (FPMQA) for community detection in online social network," *Knowledge-Based Syst.*, vol. 50, pp. 246–259, 2013.
- [9] L. Huang, R. Li, H. Chen, X. Gu, K. Wen, and Y. Li, "Detecting network communities using regularized spectral clustering algorithm," *Artif. Intell. Rev.*, vol. 41, pp. 579–594, 2014.
- [10] Y. Li, K. He, K. Kloster, D. Bindel, and J. Hopcroft, "Local Spectral Clustering for Overlapping Community Detection," *ACM Trans. Knowl. Discov. Data*, vol. 12, no. 2, pp. 1–27, 2018.
- [11] Y. Zhang et al., "WOCDA: A whale optimization based community detection algorithm," *Phys. A Stat. Mech. its Appl.*, vol. 539, p. 122937, 2020.
- [12] C. Pizzuti, "A multiobjective genetic algorithm to find communities in complex networks," *IEEE Trans. Evol. Comput.*, vol. 16, no. 3, pp. 418–430, 2012.
- [13] M. Rosvall, D. Axelsson, and C. T. Bergstrom, "The map equation," *Eur. Phys. J. Spec. Top.*, vol. 178, no. 1, pp. 13–23, 2009.
- [14] M. Arasteh and S. Alizadeh, "A fast divisive community detection algorithm based on edge degree betweenness centrality," *Appl. Intell.*, vol. 49, pp. 689–702, 2019.

- [15] U. N. Raghavan, R. Albert, and S. Kumara, "Near linear time algorithm to detect community structures in large-scale networks," *Phys. Rev. E - Stat. Nonlinear, Soft Matter Phys.*, vol. 76, no. 3, pp. 1–11, 2007.
- [16] M. J. Barber and J. W. Clark, "Detecting network communities by propagating labels under constraints," *Phys. Rev. E - Stat. Nonlinear, Soft Matter Phys.*, vol. 80, no. 2, pp. 026129, 2009.
- [17] X. Liu and T. Murata, "Advanced modularity-specialized label propagation algorithm for detecting communities in networks," *Phys. A Stat. Mech. its Appl.*, vol. 389, no. 7, pp. 1493–1500, 2010.
- [18] W. Li, C. Huang, M. Wang, and X. Chen, "Stepping community detection algorithm based on label propagation and similarity," *Phys. A Stat. Mech. its Appl.*, vol. 432, pp. 145–155, 2017.
- [19] B. D. Le, H. Shen, H. Nguyen, and N. Falkner, "Improved network community detection using meta-heuristic based label propagation," *Appl. Intell.*, vol. 49, no. 4, pp. 1451–1466, 2019.
- [20] S. Yazdanparast, M. Jamalabdoahi, and T. Havens, "Linear Time Community Detection by a Novel Modularity Gain Acceleration in Label Propagation," *IEEE Trans. Big Data*, vol. 7, no. 6, pp. 961–966, 2020.
- [21] Y. Xing, F. Meng, Y. Zhou, M. Zhu, M. Shi, and G. Sun, "A node influence based label propagation algorithm for community detection in networks," *Sci. World J.*, vol. 2014, 2014.
- [22] X. K. Zhang, J. Ren, C. Song, J. Jia, and Q. Zhang, "Label propagation algorithm for community detection based on node importance and label influence," *Phys. Lett. Sect. A Gen. At. Solid State Phys.*, vol. 381, no. 33, pp. 2691–2698, 2017.
- [23] M. Tasgin and H. O. Bingol, "Community detection using boundary nodes in complex networks," *Phys. A Stat. Mech. its Appl.*, vol. 513, pp. 315–324, 2019.
- [24] T. Wang, S. Chen, X. Wang, and J. Wang, "Label propagation algorithm based on node importance," *Phys. A Stat. Mech. its Appl.*, vol. 551, pp. 124137, 2020.
- [25] Y. Zhang, Y. Liu, Q. Li, R. Jin, and C. Wen, "LILPA: A label importance based label propagation algorithm for community detection with application to core drug discovery," *Neurocomputing*, vol. 413, pp. 107–133, 2020.
- [26] H. Li, R. Zhang, Z. Zhao, and X. Liu, "Lpa-mni: An improved label propagation algorithm based on modularity and node importance for community detection," *Entropy*, vol. 23, no.5, 2021.
- [27] H. Lou, S. Li, and Y. Zhao, "Detecting community structure using label propagation with weighted coherent neighborhood propinquity," *Phys. A Stat. Mech. its Appl.*, vol. 392, no. 14, pp. 3095–3105, 2013.
- [28] X. K. Zhang, X. Tian, Y. N. Li, and C. Song, "Label propagation algorithm based on edge clustering coefficient for community detection in complex networks," *Int. J. Mod. Phys. B*, vol. 28, no.30, p. 1450216, 2014.
- [29] K. Berahmand and A. Bouyer, "A Link-Based Similarity for Improving Community Detection Based on Label Propagation Algorithm," *J. Syst. Sci. Complex.*, vol. 32, pp. 737–756, 2019.
- [30] E. Jokar and M. Mosleh, "Community detection in social networks based on improved Label Propagation Algorithm and balanced link density," *Physics Letters, Section A: General, Atomic and Solid State Physics*, vol. 383, no. 8, pp. 718–727, 2019.
- [31] H. Jiang et al., "A Robust Algorithm Based on Link Label Propagation for Identifying Functional Modules from Protein-protein Interaction Networks," *IEEE/ACM Trans. Comput. Biol. Bioinforma.*, 2020.
- [32] P. Z. Li, L. Huang, C. D. Wang, J. H. Lai, and D. Huang, "Community detection by motif-aware label propagation," *ACM Trans. Knowl. Discov. Data*, vol. 14, no. 2, pp. 1–19, 2020, doi: 10.1145/3378537.
- [33] R. Hosseini and A. Rezvani, "ANTLP: Ant-based label propagation algorithm for community detection in social networks," *CAAI Trans. Intell. Technol.*, vol. 7, pp. 10, 2020.
- [34] K. Berahmand, S. Haghani, M. Rostami, and Y. Li, "A new attributed graph clustering by using label propagation in complex networks," *Journal of King Saud University - Computer and Information Sciences*, vol. 34, no. 5, pp. 1869–1883, 2020.
- [35] T. Zhou, L. Lü, and Y. C. Zhang, "Predicting missing links via local information," *Eur. Phys. J. B*, vol. 71, pp. 623–630, 2009.
- [36] J. Wang, M. Li, H. Wang, and Y. Pan, "Identification of essential proteins based on edge clustering coefficient," *IEEE/ACM Trans. Comput. Biol. Bioinforma.*, vol. 9, no. 4, pp. 1070–1080, 2012.
- [37] W. W. Zachary, "An Information Flow Model for Conflict and Fission in Small Groups," *J. Anthropol. Res.*, vol. 33, no. 4, pp. 452–473, 1977.
- [38] D. Lusseau, K. Schneider, O. J. Boisseau, P. Haase, E. Slooten, and S. M. Dawson, "The bottlenose dolphin community of Doubtful Sound features a large proportion of long-lasting associations," *Behav. Ecol. Sociobiol.*, vol. 54, pp. 396–405, 2003.
- [39] Krebs, V.: A network of co-purchased books about US politics. October, vol. 20, no. 1, pp. 0–03, 2008.
- [40] M. E. J. Newman, "Finding community structure in networks using the eigenvectors of matrices," *Phys. Rev. E - Stat. Nonlinear, Soft Matter Phys.*, vol. 74, no. 3, pp. 36104–36123, 2006.
- [41] M. E. J. Newman, "Network data", 2013, <http://www-personal.umich.edu/mejn/netdata>.
- [42] J. Yang and J. Leskovec, "Defining and evaluating network communities based on ground-truth," *Knowl. Inf. Syst.*, vol. 42, pp. 181–213, 2015.
- [43] A. Lancichinetti, S. Fortunato, and F. Radicchi, "Benchmark graphs for testing community detection algorithms," *Phys. Rev. E - Stat. Nonlinear, Soft Matter Phys.*, vol. 78, no.4, pp. 46110–46115, 2008.
- [44] M. E. J. Newman, "Modularity and community structure in networks," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 103, no. 23, pp. 8577–8582, 2006.
- [45] L. Danon, A. Díaz-Guilera, J. Duch, and A. Arenas, "Comparing community structure identification," *J. Stat. Mech. Theory Exp.*, vol. 2005, no.09, pp. 9008, 2005.

Natural Language Processing for the Analysis Sentiment using a LSTM Model

Achraf BERRAJAA

Euromed Research Center, Euromed University of Fes, Morocco

Abstract—Over the past decade, social networks have revolutionised the communication between organisations and their customers, and the data provided by customers on social network platforms is having an increasingly important impact on how organisations collect and analyse this data to make better decisions. We have prepared a new dataset that will allow the scientific community to estimate and evaluate new models using nearly the same conditions. Moreover, this dataset represents a recent and interesting sample for the proposed machine learning models to correctly identify the topics or points on which the company should focus to improve customer satisfaction and better meet their needs. Therefore, we have proposed a recurrent neural network (RNN) with Long short-term memory (LSTM) that we will run in the cloud to predict sentiment analysis. The objective is also to define systems capable of extracting subjective information from natural language texts, such as feelings and opinions, with the aim of creating structured knowledge that can be used by a decision support system or a decision maker for better customer management. The proposed neural network has been trained on the proposed dataset which contains 50 000 customer observations. The performance of the proposed architecture is very important as the success rate is 96%.

Keywords—Artificial intelligence; NLP; RNN; LSTM; customer relationship management

I. INTRODUCTION

In recent years, social networks have revolutionised communication between organisations and their customers. The data provided by customers on social media platforms is having an increasing impact on how organisations collect and analyse this data to make better decisions. Natural language processing (NLP) is one of the most promising ways to process data and text from social networks. Developing powerful methods and models to extract relevant information from large amounts of data from multiple sources and languages is a complex challenge. It can be overcome when a powerful pipeline is built to transform this raw data into useful information that we can use.

Information extraction [27], classification and grouping methods [25] are one of the main approaches to leverage raw textual data and transform it into valuable information. In the field of data processing, a robust pipeline is needed to clean this textual data and make it ready for use by different models. To this end, our objective is to propose a data preparation pipeline ranging from data processing and cleaning to digital representation of textual data. The prepared data is intended to be used as training data for machine learning models. A deep learning model is also implemented for sentiment analysis, defined as a system for extracting subjective information from natural language texts, such as feelings and opinions, to create

structured knowledge that will be used by a decision support system or a decision maker.

As far as structure is concerned, this paper is organised as follows: Section 2, presents a literature review on the works related to the different technologies used for natural language processing and their fields of use, namely linguistics and artificial intelligence. Section 3 details the structure of our data preparation pipeline which consists of six modules: data cleaning, tokenization, data normalization, stemming and lemmatization, token categorization and finally data representation. Section 4 is reserved for the creation of machine learning models that are able to classify the observations in our data by assigning them a star score ranging from one to five (Very Satisfied, Satisfied, Fair, Dissatisfied and Very Dissatisfied). In Section 5, a Recurrent Neural Network (RNN) with Long Short-Term Memory cells (LSTM) is implemented for the sentiment analysis of our customers. Section 6 is devoted to the digital experiments. We also discuss and comment on the results obtained and show the effectiveness of our models. Finally, a conclusion that summarises the study and gives some potential perspectives for the developed approach to improve the current results.

II. RELATED WORK

Social networking is a phenomenon that has recently developed worldwide and has rapidly attracted billions of users. The main reason for this phenomenon is the ability of online social networks to provide a platform for users for better communication, as explained in the work of [16]. This form of electronic communication through social networking platforms allows customers to generate their content and share it in different forms, usually in text form. This content is very valuable to the organisations involved. Therefore, automatic natural language processing is formed as an emerging area of research and development [19].

Natural Language Processing can be defined as a field of study using computer science, artificial intelligence and linguistic concepts to analyse natural language. In other words, NLP is a set of tools used to derive meaningful information from textual data and generally used to obtain knowledge and decision support by processing textual data present in web pages, documents, customer reviews [20].

As mentioned earlier, linguistics is an essential part of natural language processing and can be defined as the scientific study of the structure and development of language with particular emphasis on grammar, semantics and phonetics. In other words, linguistics is primarily concerned with the design and evaluation of language rules. If we read this definition

carefully, we realise that natural language is controlled by a set of rules such as grammar and semantics. These rules will be a key factor in enabling the machine to understand the textual data and to process it. The author in [13] present a study that summarises the linguistic research techniques used in automating the analysis of the linguistic structure of language and, at the same time, sheds light on the development of core technologies such as speech recognition, speech synthesis and machine translation using artificial intelligence.

Artificial intelligence (AI) is a branch of computer science that aims to propose and construct systems able of performing tasks that require human intelligence. In other words, any algorithm or computer technique capable of performing sophisticated tasks such as driving a car or diagnosing a disease can be classified as artificial intelligence (for more information, see [24]). Machine learning is a sub-field of artificial intelligence that deals with the development of algorithms capable of learning to perform tasks automatically on the basis of a large number of examples, without being pre-programmed to do so in the case of supervised learning, or with the creation of clusters and grouping of the most similar observations in the case of unsupervised learning. We cite the book *Machine learning algorithms* [8] for more information on machine learning algorithms. On the other hand, Deep Learning refers to the branch of machine learning based on neural network architectures. In [17], the authors summarise the basic principles of deep learning and machine learning to provide a general understanding of the methodical foundations of current intelligent systems. The development of NLP applications relies heavily on machine learning and deep learning methods. Therefore, both disciplines play an important role in the development of this project.

In a natural language processing project, the data is nothing more than text - unstructured data produced by people to be understood by others. Nevertheless, this unstructured data contains patterns or indications that allow it to be analysed by a computer. To put it differently, although text is unstructured data, it is not disordered. On the contrary, text is governed by linguistic properties that enable communication [10]. Thus, our contribution will be to find patterns in the text, and analyse them for better decision making. This approach is similar to that of [20] who proposed a framework for big data analytics in commercial social networks for sentiment analysis and fake review detection for marketing decision making. Furthermore, when working with natural language, we frequently encounter the concepts of syntax and semantics. The syntax of a language refers to the rules that govern the way in which linguistic elements are put together to form phrases, clauses, and sentences. It should be noted, however, that the syntactic rules of a natural language are not as strict as those of computer languages. This means that it is essential that a sentence follows the basic syntactic rules, so that it can be used to enable the machine to process textual data.

Natural language processing is performed on text data ranging from a few words entered by the user for an Internet search to multiple documents to be analysed and from which information needs to be extracted for better decision making. Thus, natural language processing is used in a variety of situations to solve many different types of problems:

- Searching for a string of characters: automatic pro-

cessing of natural language makes it possible to identify specific elements in the text. For example, it can be used to find the occurrence of a word or more generally a string of characters in a document. As an illustration, in the work of [21], a string recognition method based on a lexical search method was proposed. In this method, the string is identified by searching a sequence of segment patterns matching the string in a lexicon.

- Entity recognition: this involves extracting names of places, people, organisations and products from the text. As an application, [26] proposed an NLP pipeline from part-of-speech tagging, through chunking, to named entity recognition of Twitter.
- Sentiment analysis: This technique is used to determine people's feelings and attitudes towards a product or service. It is useful for providing feedback on how a product or service has been perceived. Companies in all sectors are analysing their social network data streams to better understand their customers' opinions. The main challenge is to extract reliable textual reviews from consumers and use them automatically to evaluate the best products or brands. [20] proposed a framework to automatically analyse reviews. Sentiment analysis was used to analyse online reviews on Amazon. Similarly, in the work of Kumar and Sebastian (2012), a hybrid method is proposed, which uses corpus-based and dictionary-based algorithms to define the semantic orientation of opinion words in tweets. In [3], a new method of approaching sentiment classification is proposed based on Twitter data.
- Search engine: Making extensive use of natural language processing for a variety of tasks, such as query understanding, query expansion, question answering, information retrieval, ranking and clustering of results. For example, in [1], a search engine specific to scientific research is proposed, which first collects the information disseminated on the web in the sites of academic institutions and in the personal homepages of researchers. Then, after intensive text processing, it summarises the information in an enriched and user-friendly presentation oriented towards non-expert users. A question answering system has been proposed by [23], which consists of automatically answering a question posed by a human in natural language using a pre-structured database or a collection of natural language documents.
- Electronic messaging: Email platforms use natural language processing to provide a series of features, such as spam classification, inbox priority, auto-completion. As an example for spam classification, [18] proposed a comprehensive spam classification system based on semantic text classification using NLP and URL-based filtering. Similarly, [12] compared various ML algorithms and a convolutional neural network was studied with the objective of creating a powerful and efficient model for correctly classifying emails.

The above use cases are just a sample. NLP is increasingly used in several other applications, and new applications of NLP

are emerging very rapidly. For instance, program synthesis is an ambitious and new field that consists in generating a source code (programming code) from a natural description. The author in [6] proposed an RNN with LSTM cells that generates java source code from a description expressed in natural language.

Several factors make the process of processing text data difficult. The existence of several languages, each of them having different rules [10] is an example of this difficulty. To illustrate, words can be ambiguous and their meaning may depend on their context. A word can also express an action, a feeling, a name or something very different. As textual data belong to the family of unstructured data, it is necessary to go through a set of data preparation operations to make them useful and usable. Despite these difficulties, natural language processing is able to perform complicated tasks in an adequate way and to bring added value in many domains. For example, sentiment analysis can be performed on customer reviews, which can identify potential problems and anomalies with a certain product or service and improve it. In the following, a robust pipeline is used to clean up the textual data and make it ready for use by the various AI models. Postal services have been considered as an application. A pipeline is proposed for data preparation, ranging from data processing and cleaning to digital representation of textual data.

III. DATASET, EXPLORATORY DATA ANALYSIS AND PIPELINE

One of the most important applications where customer service needs to be improved is the postal service. Thus, our goal is to project this application to improve this service.

A. The Dataset

The first step in the development process of any NLP classification system is to collect data relevant to the proposed problem. An initial idea that may arise is to use structured datasets found on sites such as kaggle, but the problem with this approach is that these datasets deal with topics that are not relevant to the business knowledge related to the services offered by the Post. This will result in poor performance for our model. Therefore, we will need to think about retrieving a meaningful and real training dataset for the Post. One way to build a meaningful dataset is to collect reviews from organizations that operate in very similar, if not identical, domains to the Post using a scrapping method. The idea of choosing several organizations instead of one is on the one hand to have a large volume of data and on the other hand, because very often an organization tends to have mainly positive opinions and at the same time some negative opinions (having diffuse opinions).

In order to build the most accurate and robust model possible, it is necessary to have a fairly generalized set of data and to touch as much as possible all categories of reviews. To do so, several datasets from different countries (United States Postal Service, Canada Post and French Post) have been collected. Fig. 1 shows the percentage of observations of the data with a score ranging from 1 to 5 (note that we collect and scrap the data in open source).



Fig. 1. The Percentage of Observations with a Score Ranging from 1 to 5.

Due to the fact that Post France notices are in French. The Google API was used to translate the data observations to obtain standardized observations in one language, English.

B. Exploratory Data Analysis

The idea behind this part (topology of text data observations) is to get an idea of how the data was constructed and if there is some pattern in the way reviews are written by customers, i.e. words used by customers, length of comments, etc. This will help us choose the Hyper-parameters for the Machine Learning / Deep Learning models. We start by visualizing the length distribution of the data observations in Fig. 2. To do this, we can calculate the length of each review and then visualize them using a histogram (Fig. 3).

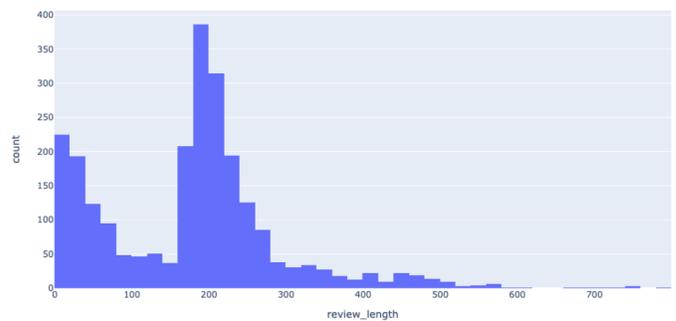


Fig. 2. Histogram of the Length of Characters per Observations.

We can say that most comments are between 180 and 219 characters long, and as the number of characters' increases, these comments become rarer. On average, a comment contains about 176 characters with a median of 190.

We can see that most of the customer reviews are about 30 to 39 words long, with an average of 32 words per review and a median of 34 and a maximum of 151 words.

This will allow us to have a first interaction with the textual data we have in order to extract indications and a first general

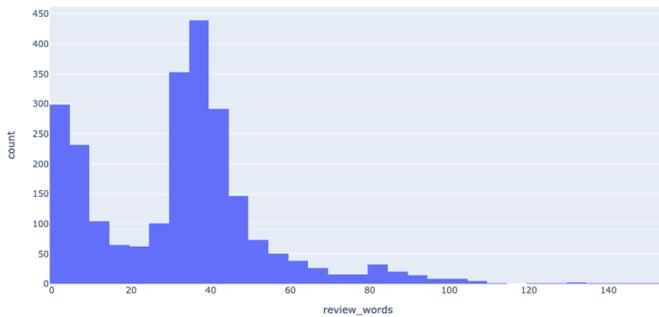


Fig. 3. Histogram Representing the Number of Words per Observation.

view of the available data.

C. Data Preparation Pipeline

At this point, the goal is to clean the data to make it more usable. This task consists of a series of steps to process, clean and normalize the textual data into a form that contains much less noise. The general idea is to remove unnecessary elements that may interfere with decision making or handicap the learning process.

The structure of our data preparation pipeline consists of six modules: data cleaning, tokenization, data normalization, stemming and lemmatization, token categorization and finally data representation.

1) *Data Cleaning*: Regular expressions allow us to create string patterns and use them to find or substitute specific strings in textual data. Python offers a rich module named "re" for creating and using regular expressions [15].

The idea behind this step is to remove characters and symbols that are usually non-alphanumeric characters, which add extra noise to text data. Regular expressions are used to detect non-alphanumeric characters and remove them afterwards.

2) *Tokenization*: Tokenization is the process of dividing a text into "tokens", a token being a significant unit of text, very often a word, that we wish to use to perform an analysis. This step is essential in any natural language processing project, especially with ML models. The importance of this task will become even more apparent when representing textual data in digital format.

We performed two types of tokenization: a) Sentence tokenization is the process of breaking up textual data into sentences. A user review may or may not contain several touching sentences about the same topic, the idea is to try to structure the observation into meaningful sentences. In order to perform tokenization, we will use the *nlTK* library using the *nlTK.sent_tokenize()* function, which has been pre-trained and gives impressive results. b) Word tokenization is the process of breaking a sentence down into its component parts. A sentence is a collection of words, and with tokenization, we basically break the sentence down into a list of words.

3) *Data Standardization*: In order to make the NLP system more robust, it is essential to go through this step where we will normalize the data by making all the tokens lowercase or uppercase. We can see the value of this step in case we are

looking for patterns in our data, or we are trying to match a certain word, but the main usefulness of this step is seen in the Feature Engineering step, where we will represent our textual data in a digital format. If this step is omitted during the data cleaning process, we will get a very high dimensional vector representation that will handicap the learning models considerably. We apply this process to all data observations using the *lower()* method.

4) *Stemming and Lemmatization*: Stemming and lemmatization ensure that the different forms a word can take, i.e. plurality, gender, cardinality, tense, etc., are treated as single vector components, which reduces the feature space during digital representation and makes the models more efficient during learning.

Stemming is important in the sense that it helps normalize words to their basic root, which facilitates many applications such as classification or any other ML algorithm. This task is performed via the "SnowballStemmer" where we specify an input language which for our project is obviously English, the stemmer removes suffixes and endings from the words and transforms them to their basic form.

A lemma is the basic form of a word. In other words, it is the form in which the word would appear if it were listed in a dictionary. The result of stemming is not always a correct word, but the lemma of a word will always be present in the dictionary. In the application of lemmatization, an additional step is involved where the formed root is compared to the information provided by the dictionary and if and only if the lemma is present in the dictionary it can be taken into consideration. This makes undoubtedly this step much more complex and computationally demanding. The research proposed in [5] is based on comparing the accuracy performance of document retrieval based on language modeling technology, especially stemming and lemmatization. In addition, a baseline ranking algorithm was used to compare the two technologies.

D. Categorization of Tokens

1) *Part-of-Speech Tagging*: Part-of-speech (POS) is a process that determines the grammatical category of each token in the textual data. It labels nouns, verbs, adjectives, adverbs, interjections, conjunctions, singular nouns, plural nouns, proper nouns, etc. All this is done by models trained on datasets containing tokens and the associated tag. This is a crucial step for many NLP applications because by identifying the POS of a word, one can infer its meaning in context, which is used by many machine learning models to better classify textual data. The same token can sometimes be a noun, a verb or adverb. This is where Part-of-Speech tagging becomes of great importance since it allows us to categorize tokens as precisely as possible based on a statistical approach. To add, we use the POS model of the *Spacy* library to categorize the tokens of textual observations.

The same word can have different interpretations. This is why grammatical structure is important. The *NLP spacy* library uses models that have been pre-trained to best predict the usage of a word in a data observation. When analyzing customer reviews, we want to know what actions customers have performed or undergone in order to get to know them

better and this is where POS tagging comes in, allowing us to easily retrieve all actions performed by the customer.

2) *Named Entity Recognition*: When examining textual data observations, we tend to first identify the key actors in the observation, such as people, places, and organizations. This classification helps us break up data observations into entities and make sense of the semantics of the data observation. The Named Entity Recognition (NER) mimics the same behavior and is used to classify entities (tokens).

Named Entity Recognition (NER) is a process that detects names in textual data, as well as dates, monetary amounts and other types of entities. NER tools often focus on three types of categories: Person, Organization, and Location, drawing on models that have been pre-trained. The Named Entity Recognition process consists of two sub-tasks: entity detection (the token) and entity type determination based on statistical methods, very often supervised models. NER models rely on several parameters, one of which is the POS (part-of-speech-tagging) of the token. It is at this point that we can see a clear relationship between the POS and NER process.

E. Data Representation

Feature engineering is an important step for any machine learning problem. No matter how good the learning algorithm used, if we introduce bad features, we will get bad results.

At this point, we only have cleaned textual data, but we cannot feed it to a machine learning model. Therefore, it is essential to find a way to represent this data so that we can process it with a given model. In other words, we need to transform our textual data into digital form so that it can be passed to ML algorithms.

In this task, the objective was to work on different methods of representing text as vectors in order to choose the best one for our learning models. Several data representations have been used in the field of data science, such as Bag-of-Words [28], TF-IDF [2], Word Embeddings [4] and Word2vec [11]. We chose to use TF-IDF because of its compatibility with this problem.

TF-IDF aims to quantify the importance of a given word relative to other words for each observation in our dataset. The problem with the bag-of-words is that, since the feature vectors are based on the frequency of tokens, some terms may appear frequently in all observations in our dataset and tend to mask the importance of other words in the feature set. In particular, words that do not appear as frequently, but may be more interesting and important as characteristics. A simple way to think about the TF-IDF process is as follows: if a word m appears many times in an observation, but does not appear much in the rest of the observations in the dataset, then word m should have high importance for the observation in question. The importance of m should increase in proportion to its frequency in the observation in question, but at the same time its importance should decrease in proportion to the frequency of the word in the other observations.

Mathematically, TF-IDF is the product of two metrics. TF measures the frequency of occurrence of a word in an observation. Since different observations in the dataset may be of different lengths, a term may appear more often in a

long observation than in a short observation. To normalize this, we divide the number of occurrences by the length of the observation. The IDF measures the importance of a word in the entire dataset. We calculate it by dividing the total number of observations in our dataset by the number of observations containing the term. We do this for each term and then apply a logarithmic scale to the result.

The TF-IDF score is a product of these two metrics. Thus, the TF-IDF score = TF * IDF. And can be represented as follows:

$$W_{i,j} = tf_{i,j} \times \text{Log}\left(\frac{N}{df_i}\right)$$

We apply this same model to our dataset and this numerical representation will be used several times over the prediction models.

IV. CONSTRUCTION OF A REVIEW SCORE PREDICTION MODEL

At this stage, we explored our data, and were interested in the syntax, structure and semantics of the observations in our dataset. We also cleaned our data; then the data was represented in a digital format using the TF-IDF algorithm. The next step is to build models that allow us to take advantage of all the steps seen in the previous tasks. In this step, the task was to create a model that predicts customer ratings with the highest possible accuracy. To do this, a first approach was to understand the objective and how to go about it. The problem addressed, which is the prediction of the star score of reviews, is a classification task that consists of classifying reviews into five categories ranging from a score of one to five based on the specific properties or attributes of each review. The classification of textual data is one of the most complex tasks in automatic natural language processing due to the many factors involved, namely the quality of the data cleaning, the algorithm used to represent the textual data, the quality of the data used to train the model, the prediction model used and its parameters, etc. As we can see, there are many factors and variables and the objective is to optimize each of these factors to obtain the best possible result.

What customers think is important information, it is very valuable knowledge for the organization concerned in the sense that it gives a very clear idea of the quality of a certain product or service and how the public perceives it. While some platforms allow users to give a star rating, most social networks do not offer this possibility of rating on a scale of one to five stars. The objective of this task will be to create a machine learning model that will be able to rank our data observations by assigning a star rating from 1 to 5 (Very Satisfied, Satisfied, Average, Dissatisfied, and Very Dissatisfied).

Once we have a numerical representation of our training dataset, we need to use classification algorithms, which are nothing more than supervised learning algorithms used to classify data observations into different categories. The goal at this point is to train classification models on our training dataset. The classification algorithm will identify patterns based on the characteristics of our training data observations and their corresponding labels, and the result of this identification will constitute our score prediction model. The models are expected to be generalized enough to predict classes for new

data observations (without labels). Thus, having a generalized dataset is essential, as has already been pointed out.

Several learning algorithms were used for this task: SVM, KNN, Decision tree, Random forests and Logistic regression.

The model evaluation is an estimate that can be used to indicate how well you think the algorithm can actually perform. It is not a guarantee of performance since we are talking about a statistical estimate. Once we have estimated the performance of our algorithm, we can re-train the final algorithm on the training data set and prepare it for operational use. The performance of classification models is based on their ability to predict the outcome of new observations. This performance is measured against a test data set, which consists of observations that were not used to train the model. This textual dataset is a real post observation set that represents a real subset (which will be the test dataset of our model). This test dataset contains observations and their corresponding labels. The digital representation of the test dataset is fed into the already trained model and predictions are obtained for each observation. These predictions are then compared to the actual labels to determine the quality or accuracy of the model prediction.

In the result section we will detail the percentage of success of each model. Based on these results, we can observe that most of the models have a good performance, the multinomial logistic regression having the best performance and, on the other hand, the decision tree having the worst predictions.

V. CONSTRUCTION OF A SENTIMENT ANALYSIS MODEL

Sentiment analysis is one of the most active research areas in natural language processing. Sentiment analysis is a field of study that analyzes people's opinions, feelings, evaluations, attitudes and emotions through natural language.

The aim of sentiment analysis is to define a system that can extract subjective information (such as opinions and feelings) from natural language to create structured knowledge that can be used by decision support systems or decision makers. Nowadays, with the advent of social networks, sentiment analysis has gained in value. Their wide diffusion and their role in modern society represent one of the most interesting developments in recent years, attracting the interest of organizations and companies. Not so long ago, sentiment analysis was almost non-existent. Opinions were collected through surveys rather than through observation of textual data because computers were not capable of storing and processing large amounts of data, and there were no algorithms for extracting knowledge from written language.

The explosion of sentiment-laden content on the Internet, the increase in computational power, and advances in data mining techniques have made sentiment analysis a burgeoning research area and a crucial business sector. In this classification, we have implemented a recurrent neural network (RNN) with LSTM cells, where we will apply it on a digital representation of the data represented by One Hot Encoding [9], in order to classify the observations of the Post's customers into Very Satisfied, Satisfied, Average, Dissatisfied, and Very Dissatisfied opinions.

Recurrent neural networks are more effective than traditional neural networks and machine learning algorithms at retaining information from the previous event. The recurrent neural network involves a combination of loop networks. The loop network allows the information to persist. Each network in the loop takes the input and information from the previous network, performs the specified operation and produces an output, and at the same time passes the information to the next network. Some applications only need the most recent information, while other applications may need more information from the past, such as NLP (the meaning of related words). As the gap between the required prior information and the point of application has increased to a large extent, the learning of simple RNN lags behind. But fortunately, long-term memory networks [14], a special form of RNN, can learn such scenarios.

Long-term memory (LSTM) is an alternative architecture proposed in [14]: the traditional architecture of a Recurrent Neural Network (RNN) that is based on a simple activation function is modified in such a way that the vanishing gradient problem is explicitly avoided, while the learning method remains unchangeable. For more information on this architecture [6], [7]. But what are the strengths of LSTM? why the LSTM will be effective to solve the steel continuous casting problem?

A LSTM neuron network is made up of several cell that have not just one activation function but rather three that are represented as an input gate, a forget gate and an output gate. Each cell remembers the state of the problem treat in several time intervals, and the three gates regulate the flow of information in and out of the cell. The LSTM network is very suitable for classification, processing and prediction based on time series data, because there may be lags of unknown duration between important events in the time series. This is what is needed to understand a series of words in a sentence. Also as we explained, LSTM was developed to deal with the explosion and disappearance of gradients that may be encountered when training traditional RNNs.

A. The Proposed LSTM Architecture and Hyperparameter

To define a neural network, it is necessary to establish parameters, such as training data, type of neural network, number of layers, connections, activation functions, propagation rules, etc.

There are several ways to train a neural network to produce a specific output for a specific input. In the current training method (Forward / Backward Propagation), we have error propagation, which involves adjusting the network based on each neuron's contribution to the error, and each neuron adapts the weights by using the gradient of descent. Genetic algorithms are also used to train neural networks [22]. By training these networks on a dataset with known correct outputs, the network will be able to return an approximate result for new data (not seen in the training stage). We trained our LSTM with data from 50,000 use cases (customer observations).

Regarding the architecture, there are several architectures proposed in the neural network language model with some differences between them. The proposed architecture for our application area consists of some basic principles, such as:

The input of our LSTM (the input sequence) is represented by a sentence, it is coded by the code from 1 to K, where K is the length of the client's observation. This involves the use of One Hot Encoding [9] to list individual words. The sequences are generated in 128 batches with 5% diversity (with the goal of avoiding overfitting).

The network topology is represented as follows: An input layer that takes as input the digital representation of the data observations. Five hidden layers have been implemented, each with 256 LSTM units, and each uses the "relu" activation function. For the five hidden layers, a normalization step is added via the Batch Normalization technique, which improves the performance and stability of neural networks. The main idea is to normalize the inputs to each layer so that they have an average output activation of zero and a standard deviation of 1. The regularization is used in the network in the form of an exclusion layer. This form of regularization prevents overfitting of the model. The output layer is composed of five neurons, as we want to perform a five-class classification with a "softmax" activation function that will be used to represent a probability distribution to predict customer sentiment. As shown, the neural network trains by adjusting the weights by comparing the results predicted by the neural network and the actual label of the observations in the dataset.

Now that we have built the model, we will have to train it on the digital representation of the training data. To do so, we will have to define a cost function which measures the difference between the results predicted by our model and the real results. As this is a classification case, we tested several functions but the experimental results show that the categorical function "cross-entropy" is the best for this application domain. So we have a neural network, a cost function to be minimized. To minimize the cost function, we used the gradient descent algorithm and exactly the Adam optimizer which is an optimization algorithm to minimize convex functions. As a learning standard, the used precision of the Adam optimizer is 0.001.

Finally, we define precision as an evaluation metric for our neural network. The following algorithm summarizes the main steps:

VI. EXPERIMENTAL TESTS

In order to evaluate the performance of our models for the two treated part, this part puts forward the performance of the models taking into account the quality of the solution. Note that all the experiments were performed on the google colab under GPU.

A. Assessment of Classification Models

The evaluation metrics used for classifying textual data (observations of data with a star score ranging from 1 to 5) are generally as follows:

Accuracy: is the sum of true negatives and true positives divided by the total number of observations.

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

Algorithm 1 : The Proposed LSTM for Better Customer Relationship Management.

- Data preparation pipeline (see section 3).
 - Creation of the LSTM architecture : Creating eight layers, one input layer, five hidden layers each with 256 LSTM units, dropout layer for regularization and final output layer consists of five neurons and "softmax" activation function is used to predict the customer feelings (Very Satisfied, Satisfied, Fair, Dissatisfied and Very Dissatisfied).
 - According to the unified law, the weights are initialized randomly.
 - Codage : list all customer's observations. The One Hot Encoding is used to represent each observation and coded it.
 - The categorical "cross-entropy" is used as cost function.
- while** index \leq Max_iter **do**
1. Five percent of diversity is generated in each batch.
 2. Measuring the difference between the results predicted by our model and the actual results, using cost function.
 3. Adam optimizers is used with a precision of 0,001 in order to minimize the cost function.
 4. Update the weights.
- end while**

Where TN = True Negatives, TP = True Positives, FN = False Negatives and FP = False Positives.

Precision: is the number of real positives on the set of positive cases predicted by the model.

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

Recall: allows the performance of the model to be evaluated from the point of view of the positive class. It indicates the percentage of actual positive cases that the model is able to predict correctly out of the total number of positive cases.

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

F1 Score: a combination of "Precision" and "recall".

we can observe that most of the models have a good performance (a cause of the proposed data processing pipeline), the multinomial logistic regression having the best performance and the decision tree having the worst performance (Table I).

TABLE I. METRICS ASSOCIATED WITH THE APPLICATION OF MODELS ON THE DATA REPRESENTED BY TF-IDF.

Algorithms	Accuracy	Precision	Recall	F1 Score
Random Forst	0.8487	0.7203	0.8487	0.7792
Decision Tree	0.6513	0.7971	0.6513	0.7792
Multinomial Logistic Regression	0.8553	0.8017	0.8553	0.8257
SVM	0.8289	0.7952	0.8289	0.8079

Learning models are parameterized so that their behavior can be tailored to a given problem. Models can have many parameters, and finding the best combination of parameters

can be treated as a search problem. To treat a search problem, we can use different search strategies to find a parameter or a robust set of parameters for an algorithm on a given problem, namely grid search.

Grid search is a parameter tuning approach that allows you to systematically build and evaluate a model for each combination of specified algorithm parameters. To perform Grid Search tuning to find the best parameters, we use the *GridSearchCV* class, but it is important to remember that this task consumes a lot of computing resources and takes a long time to return results.

We specify a few parameters for our model to experiment with, we can choose any values that make sense and the grid search model will return the best set of parameters.

Once our parameters are ready, all we have to do is initialize a gridsearch object and use it. To do this, we execute the parameters we have prepared and also specify the model and the metric that will be used to choose the best model. Once the gridsearch model is ready, we apply it to our data. We repeat this same step for the rest of the models and visualize the results as shown in Fig. 4. We can see that the multinomial logistic regression has the best performance.

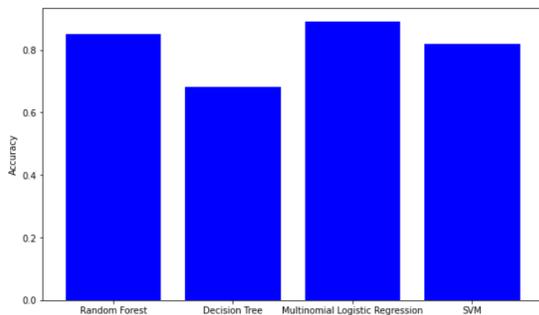


Fig. 4. Diagram of the Accuracy Metric of the Models after Applying the Gridsearch.

B. LSTM Network Assessment

To evaluate the RNN model with LSTM cells, we divided our data set (50,000 observations) into two sets, one for training and one for validation, and used cross-validation for training.

1) *The Learning and Validation of LSTM Model:* The learning rate gives an idea on the improvement of the quality of learning on a model. In Fig. 5, we present a graph that represents the learning rate (96%) and the validation rate (89%) of the proposed LSTM.

As can be seen from the model validation accuracy visualization, our RNN model with LSTM cells performed well and the fact that there is not a large discrepancy between the learning accuracy and the validation accuracy allows us to conclude that we do not have an overfitting (Fig. 6). We have also to mention that the learning rate (96%) and the validation rate (89%).

By comparing the accuracy parameters of LSTM with those of machine learning models, we can conclude that the learning rate is higher for LSTM, which will lead to good

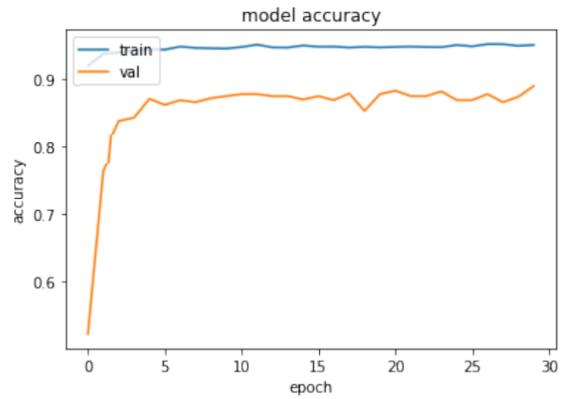


Fig. 5. Development of the Training and Validation Score per Epoch.

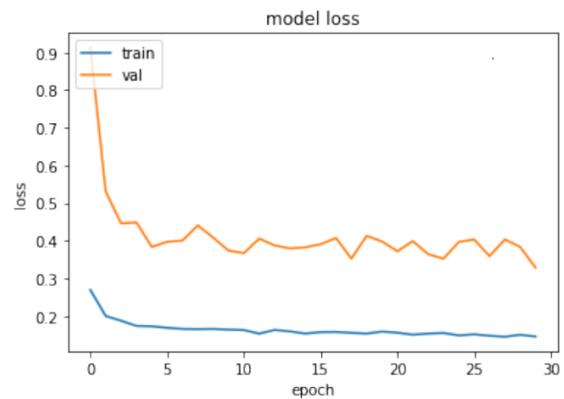


Fig. 6. Convergence of the Cost Function per Epoch

results. To test our model, we tested it with new and real data (not used in training). For this, we retrieved 1000 new observations (customer reviews with different classes) and compared the predictions with customer satisfaction (target provided by customers). The following confusion matrix (Fig. 7) shows the results of the test where the accuracy mitrics is equal to 94%.

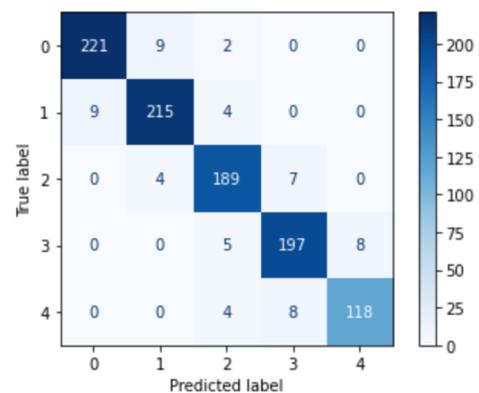


Fig. 7. The Confusion Matrix of the Test Set.

VII. CONCLUSION

In this paper, we have implemented many tasks to propose a data preparation pipeline ranging from data processing and cleaning to digital representation of textual data. We also performed Feature Engineering by digitally representing the text data using different algorithms.

The training data was collected from several companies operating in the same field as the position and the data set had to be balanced as much as possible, which was a challenge. Once the training database was formed, different models were prepared (SVM, KNN, decision tree, random forests, and logistic regression), in order to find the most effective model to rank the observations in our data by assigning them a star score ranging from one to five. After developing the score prediction model, we turned to developing a sentiment analysis model using RNN with LSTM. The performance of the proposed LSTM is very interesting such that the success percentage is 96%.

One of the future works we plan to develop is to generalize the approach on a GPU's cluster platform in order to process more complex documents and not only sentences.

ACKNOWLEDGMENT

The author would like to thank the Editor-in-chief and anonymous reviewers for their comments and suggestions to improve the quality of the paper.

REFERENCES

- [1] Armentano M. G., Godoy D., Campo M. and Amandi, A. NLP-based faceted search: Experience in the development of a science and technology search engine. *Expert systems with applications*, 41(6), 2886-2896, (2014).
- [2] Bafna P., Pramod D. and Vaidya A. Document clustering: TF-IDF approach. In *2016 International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT)*, IEEE, 61-66, (2016).
- [3] Bagui S., Wilber C. and Ren K. Analysis of Political Sentiment From Twitter Data. *Natural Language Processing Research*, 1(1-2), 23-33, (2020).
- [4] Bakarov A. A survey of word embeddings evaluation methods. *arXiv preprint arXiv:1801.09536*, (2018).
- [5] Balakrishnan V. and Lloyd-Yemoh E. Stemming and lemmatization: a comparison of retrieval performances, (2014).
- [6] Berrajaa A. and Ettifouri E. H. The Recurrent Neural Network for Program Synthesis. In *International Conference on Digital Technologies and Applications*, Springer, Cham, 77-86, (2021).
- [7] Berrajaa A. Solving the Steel Continuous Casting Problem using an Artificial Intelligence Model. *International Journal of Advanced Computer Science and Applications*, vol. 12, no 12, (2021).
- [8] Bonaccorso G. Machine learning algorithms. *Packt Publishing Ltd*, (2017).
- [9] Buckman J., Roy A., Raffel C. and Goodfellow I. Thermometer encoding: One hot way to resist adversarial examples. In *International Conference on Learning Representations*, (2018).
- [10] Chowdhury G. G. Natural language processing. *Annual review of information science and technology*, 37(1), 51-89, (2003).
- [11] Church K. W. Word2Vec. *Natural Language Engineering*, 23(1), 155-162, (2017).
- [12] Eckhardt R. and Bagui S. Convolutional Neural Networks and Long Short Term Memory for Phishing Email Classification. *International Journal of Computer Science and Information Security*, 19(5), (2021).
- [13] Hirschberg J. and Manning C. D. Advances in natural language processing. *Science*, 349(6245), 261-266, (2015).
- [14] Hochreiter S. and Schmidhuber J. Long short-term memory. *Neural computation*, 9(8), 1735-1780, (1997).
- [15] Hunt J. Regular expressions in python. In *Advanced Guide to Python 3 Programming*, Springer, Cham, 257-271, (2019).
- [16] Jain A. K., Sahoo S. R. and Kaubiya J. Online social networks security and privacy: comprehensive review and analysis. *Complex and Intelligent Systems*, 1-21, (2021).
- [17] Janiesch C., Zschech P. and Heinrich K. Machine learning and deep learning. *Electronic Markets*, 1-11, (2021).
- [18] Junnarkar A., Adhikari S., Faganian J., Chimurkar P. and Karia D. E-Mail Spam Classification via Machine Learning and Natural Language Processing. In *2021 Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV)*, IEEE, 693-699, (2021).
- [19] Kang Y., Cai Z., Tan C. W., Huang Q. and Liu H. Natural language processing (NLP) in management research: A literature review. *Journal of Management Analytics*, 7(2), 139-172, (2020).
- [20] Kauffmann E., Peral J., Gil D., Ferrandez A., Sellers R. and Mora H. A framework for big data analytics in commercial social networks: A case study on sentiment analysis and fake review detection for marketing decision-making. *Industrial Marketing Management*, 90, 523-537, (2020).
- [21] Koga M., Mine R., Sako H. and Fujisawa H. Lexical search approach for character-string recognition. In *International Workshop on Document Analysis Systems*, Springer, Berlin, Heidelberg, 115-129, (1998).
- [22] Lamos-Sweeney J. D. Deep learning using genetic algorithms. *Rochester Institute of Technology*, (2012).
- [23] Lende S. P. and Raghuvanshi M. M. Question answering system on education acts using NLP techniques. In *2016 world conference on futuristic trends in research and innovation for social welfare (Startup Conclave)*, IEEE, 1-6, (2016).
- [24] McCarthy, J. What is artificial intelligence?, (2007)
- [25] Nugroho K. S., Sukmadewa A. Y. and Yudistira N. Large-scale news classification using bert language model: Spark nlp approach. In *6th International Conference on Sustainable Information Engineering and Technology*, 240-246, (2021).
- [26] Ritter A., Clark S. and Etzioni O. Named entity recognition in tweets: an experimental study. In *Proceedings of the 2011 conference on empirical methods in natural language processing*, 1524-1534, (2011).
- [27] Watkins H., Gray R., Jha A. and Nachev P. An artificial intelligence natural language processing pipeline for information extraction in neuroradiology. *arXiv preprint arXiv:2107.10021*, (2021).
- [28] Zhang Y., Jin R. and Zhou Z. H. Understanding bag-of-words model: a statistical framework. *International Journal of Machine Learning and Cybernetics*, 1(1-4), 43-52, (2010).

Alarm System using Image Processing to Prevent a Patient with Nasogastric Tube Feeding from Removing Tube

Amonrat Prasitsupparote

College of Computing
Prince of Songkla University, Thailand

Pakorn Pasitsuparoad

Faculty of Technology and Environment
Prince of Songkla University, Thailand

Abstract—A removal nasogastric (NG) tube of a patient is a critical problem especially the patients resist swallowing. To solve this problem, the conventional approach using a personal caretaker is a time-consuming and intense focus on the patient's hands. However, visual technology can decrease the intense focus of a personal caretaker by using image processing to evaluate the patient's gesture and warn the personal caretaker when the patient acts in a risky pose. This work illustrates the feasible solution to prevent a patient with nasogastric tube feeding on removing tube by applied the face detection using Haar and Fiducial markers which consist of color marker and ArUco marker. An image processing can evaluate the patient's gesture and warn the personal caretaker when the subject acts the risky pose. A Raspberry Pi 3 Model B and a Camera module with Python and Open CV package are applied to detect and evaluate the warning gestures with 648 measurements. Six detection methods to evaluate and warn when the patient on bed tries to remove a nasal feeding tube were performed and the results were analyzed. The results show that the detection method using ArUco marker is found to be a good candidate for the alarm system preventing nasogastric (NG) tube removal of a patient.

Keywords—NG tube; image processing; fiducial markers; face detection; Aruco

I. INTRODUCTION

The insertion of nasogastric (NG) tube feeding into the gastrointestinal tract (GI) is very irritating, high risk and it heavily damages the GI tract if the patients resist to swallow [1], [2], [3], [4], [5]. It is typical in the patients with dementia to resist and remove the NG tube. Therefore, a common approach to prevent the NG tube removal by the patients is to tie their hands with the bed or wear hand mittens [6], [7], [8], [9]. The mentioned methods leave many problems like bruises on the patient's wrists, wounds on the patient's finger and palms, or ankylose. The preferred interventions to prevent tube removal or dislodgement are taping the NG tube to the face, application of hand mittens and insertion of nasal loop systems (bridle) [10], [11], [12]. Unfortunately, these are the best practices for keeping NG tube in place. At present these interventions are controversial among patient and relative feeling because of the GI damage, hand restrain, skin exacerbation, diminishing autonomy and justice [13].

According to the FOOD Trials, a family of three multi-centre international randomized controlled trials to address feeding issues for acute stroke patients, the results indicate that stroke patients frequently pulled out their tubes up to 18

times per a patient [14], [15]. The tube removal interrupts their nutrition, hydration or medication. Moreover, dislodging the tube may result in feed fluid entering respiratory tract and causing respiratory tract infections[16]. A long-term study on the accidental removal of endotracheal and nasogastric tubes shows that the patients accidentally remove tubes 13.1 and 41.0 times per 1000 days, respectively [17]. In the aspect of patient numbers, the accidental removal rates are 38 patients out of 289 endotracheal patients and 151 patients out of 368 NG patients. There are many factors involving the incidental removal rate such as age, intervention methods, tube placement position and consciousness level of patients [18]. According to the review, the most used and taught securing technique is using adhesive tape because of its feasibility, convenience and fairly comfortable for the patients. The biggest drawback is the high risk of tube dislodgement or removal and correlates to complications [19]. It is obvious that there is no health system assisting endotracheal or NG patients over conventional interventions.

Recently, Internet of Things (IoT) has been deployed widely in medication, rehabilitation, and intensive caring. There are many applications for autonomous patient monitoring using image processing to evaluate patient status such as postures, facial action units and expressions, head pose variation, and extremity movements [20]. The detection relies mainly on object recognition, objection position and ambient evaluation using sensors. Another example on vision assisted healthcare system is fall detection. There are many techniques used for fall detection like multi-camera systems, monocular systems, infrared and range sensor based systems, and bio-inspired vision sensor based systems [21], [22], [23], [24]. For the monocular camera systems, the camera is mounted on the ceiling or the wall. The 3D space ellipsoid is projected into 2D image plane and the object is tracked with markers [25], [26]. The camera calibration and inverse perspective mapping are performed for the area of interest. Adopting the vision technology, the risky pose can be detected to warn the personal caretaker.

In order to evaluate the risky pose of the subjects, there are many available technologies such as gesture recognition [27], Bluetooth sensors [28], RFID sensors [29], gyroscope [30], [31], face detection [32], [33], [34], and fiducial markers [35], [36]. Each technology has advantages and drawbacks depending on the application. The gesture recognition uses a stereo camera to collect 3D images. Its accuracy and sensitivity are

very high as well as the cost and calculation power. Moreover, the device size is relatively large compared to the patient. Therefore, the installation requires huge space. The Bluetooth and RFID sensors work on calculating the distance between two sensors on the patient's body. They are fast, sensitive and accurate but the sensors carry a certain area and load, approximately 10 cm^2 and 50 g. Both sensor size and weight mostly rely on the battery. In contrast, the gyroscope gives the exact 3D position of the patient. A gyroscope sensor works on the principle of conservation of angular momentum. It works by preserving the angular momentum. Therefore, the position resolution depends on the change in angular momentum and the calibration. In contrast, the image processing method works on an acquired image and algorithms to evaluate the patient's pose. However, the image is sensitive to light condition, image resolution, perspective and background colors. It provides fast, portable and cheap solution.

A face detection using Haar cascades was proposed by Paul Viola and Michael Jones [32], [33]. It is a machine learning based approach where a cascade function is trained from a lot of positive and negative images. It is then used to detect objects in other images. The algorithm needs a lot of positive images (images of faces) and negative images (images without faces) to train the classifier. The image then tested with the trained classifier to give the result. This method requires the full subject face including two eyes, one nose and one mouth. Moreover, it is sensitive the subject face angle to the camera.

Using fiducial markers are an alternative approach to detect the patient's pose. The markers are small and light. They can be attached on the subject face and hands. The color on images can be easily detected and evaluated. The process is to set a certain range of selected color and transform the rest into black. Unfortunately, the detection of color heavily depends on light condition and the background color. Recently, Sergio Garrido and Rafael Muñoz proposed a set of marker library called ArUco to estimate the pose [35], [37], [38]. An ArUco marker is a synthetic square marker composed of black border and inner white binary matrix which determines its identifier (id). The ArUco marker is a fiducial marker designed for pose estimation in many vision applications such as robot navigation and Augmented Reality (AR). Therefore, it is very insensitive to light, image resolution, background noise, nor perspective distortion in 2D or 3D spaces [35], [36].

This work aims to propose an alarm system preventing a patient on bed try to remove a nasogastric (NG) tube which this system can be processed and executed on a small device. This system evaluates the patient's gesture and warns the personal caretaker when the subject acts the risky pose by using visual technology. Therefore a face detection and markers technique was considered in the experiment on Raspberry Pi 3 with a camera module. The contribution of this paper is as follows:

- To detect and evaluate the risky pose of the subjects using visual technology.
- To compare and evaluate detection methods using visual technology.
- To demonstrate an alarm system preventing nasogastric (NG) tube removal of a patient.

II. METHODOLOGY

The subject is equipped with markers to estimate the warning gestures of the subject. The warning gestures are defined as the distance between two markers or the subject's face detection. For the warning using distance, the alert creates when the distance between two nearest markers are less than or equal to the setting value. For the warning using face detection, the alarm creates when the camera cannot detect the subject's face in the frame. Using both warning gestures and different marker types, there are six approaches used to evaluate the warning gestures as follows:

- a) Two ArUco markers (ArucoX2): This method uses ArUco markers to find the distance between two nearest marker centers. One marker is placed on the subject's face and another one is placed on the subject's hand.
- b) Two colors markers (ColorX2): This method uses red and blue markers to find the distance between two nearest marker centers. A blue marker is placed on the subject's face and a red marker is placed on the subject's hand.
- c) Face detection (FaceOnOff): This method identifies the subject's face in the frame by using Viola-Jones framework with Haar features.
- d) An ArUco marker and face detection (FaceAruco): This method finds the distance between the center of the face and a nearest ArUco marker. The ArUco marker is placed on the subject's hand.
- e) A color and an ArUco marker (ColorAruco): This method finds the distance between a selected red marker center and the nearest ArUco marker center. The ArUco marker is placed on the subject's face and the red marker is placed on the subject's hand.
- f) A face detection and a color marker (FaceColor): This method finds the distance between a detected face center and the nearest red center. The color marker is placed on the subject's hand.

The proposed system aims to invent a small device to alarm the risky pose of a patient in the next phase of this research, therefore this work choose a Raspberry Pi with a camera module. A Raspberry Pi 3 Model B and a camera module V2 with Python and OpenCV package are applied to detect and evaluate the warning gestures. In order to find the best detection methods many test conditions are performed and scoring are given to compare the six methods. A list of test conditions is given in Table I.

Five test conditions are selected to represent the performance close to the real environment as follows:

- (i) The subject face angle with respect to the camera is performed to find the limit when the patient turns around. This measurement also represents the perspective distortion on the markers.
- (ii) The subject to camera distance represents many effects such as the perspective distortion, marker size, image resolution and the background disturbance.
- (iii) The marker size is performed to find the optimum marker size. Both ArUco and color markers are prepared as squares on white paper as shown in Fig. 1.
- (iv) The light intensity also affects the detected images mainly on the color properties like shade, tone, saturation and

TABLE I. TEST CONDITIONS FOR THE WARNING GESTURE DETECTION

Condition	Value		
(i) Subject face angle	0°	45°	90°
(ii) Subject to camera distance	30cm	50cm	100cm
(iii) Marker size	1cm ²	9cm ²	25cm ²
(iv) Light intensity	Low intensity	High intensity	
(v) Background color	Monocolor	Multicolor	

hue. Low intensity means the light in a patient's room is off, while high intensity means the light is on.

- (v) The background color on the image usually disturbs the color marker method and sometime misleads the face detection or ArUco marker detection. Therefore, the multicolor background color can provoke the disturbance.

This experiment consists of five test conditions and six approaches, thereby there are 108 measurements times six methods, 648 measurements in total.

During the test, a subject lies on a patient's bed with a camera hung above the patient's head. There are three different face angles in our experiment as shown in Fig. 2. The subject starts with the normal gesture, both hands lie parallel to the body in order to verify the false-positive results. Then, one hand is moved close to the face for the warning gesture. The markers are attached to the subject using a transparent tape. For the monocolor background, bed sheets of green and yellow are used while the multicolor background has at least 7 colors on the bed sheets and subject's clothes. The colors of markers are red and blue. The distance between the camera and subject is measured from the subject's forehead to the camera's lens. During the warning pose, if the method detects the warning, the score of 1 is registered to the performance. Otherwise, the score of 0 is registered.

According to the mentioned before test conditions, four attributes can be extracted from the performances of each method naming speed, accuracy, resolution and tolerance and reliability. A scoring system is used quantified each method. A value of 1 is given for the successful detection and 0 for fail detection. In this study, we neglect the false positive or false-negative results. There are four attributes to measure the performance as follows:

- The speed is acquired from the Frame Per Second.
- The accuracy is calculated from a summation of all test conditions (108 measurements). The accuracy shows how good is the method related the others.
- The tolerance and reliability are acquired under the hardest condition when the marker size is kept at 1 cm² (36 measurements). The score comes from the summation under this condition. It shows the robustness of the detection method.
- The resolution is calculated from the summation of test conditions when the subject face angle is kept at 0° (36 measurements). This attribute compares the effects of image resolution on each method.

The performance representation of each attribute is shown as percentage of the observables over total success cases.

III. RESULT AND DISCUSSION

A set of python codes is implemented on Raspberry Pi 3 to detect the warning gestures under several conditions. The results are recorded and analyzed according to the four attributions to estimate the best method of detection. The experiment was conducted in a patient's room which has two windows near the patient's bed. The performance representation of each attribute is shown as percentage of the observables over total success cases is shown in Table II.

All methods show no difference in the aspect of speed. The average frame per second is 2. The two colors method (ColorX2) is supposed to be the fastest method because it demands the smallest calculation time than others. Surprisingly, the results show no significant difference. The explanation lies in the Raspberry Pi 3 and the camera limits. Moreover, the two colors method (ColorX2) is very sensitive to marker size, subject to camera distance, light intensity, perspective distortion, background noise and subject's face angle which give the lowest score [39], [40]. In general, it is the worst method of detection. The performance of all methods is represented using a radar plot as shown in Fig. 3.

According to the Fig. 3, the overall best performance is assigned to the Face detection method (FaceOnOff). This method has the highest resolution regardless of the image resolution and the highest tolerance to the light intensity and background noise. On the contrary, this method requires the full subject's face including two eyes, one nose and one mouth. Therefore, it is insensitive to the subject's face angle and perspective distortion. These two effects reflect through the relatively low accuracy.

The second place in term of overall performance is the two ArUco markers method (ArucoX2). Though the markers suffer from the image resolution. The image resolution limit is still good, meaning 1cm² markers can be detected only when the distance is 30cm, not detected at the longer distances. In addition, the ArUco markers are very robust to the light intensity, perspective distortion, background noise and subject's face angle which give the highest accuracy.

Another group of detection method is the combination of color markers, ArUco markers and face detection (ColorAruco, FaceAruco, FaceColor). These three methods show relatively similar performances. Unfortunately, the combination does not give the better result than the original because they do not overcome the intrinsic problems on each detection method. They only join those drawbacks and lower the detection performances.

Six detection methods to evaluate and warn when the patient on bed try to remove a nasal feeding tube were

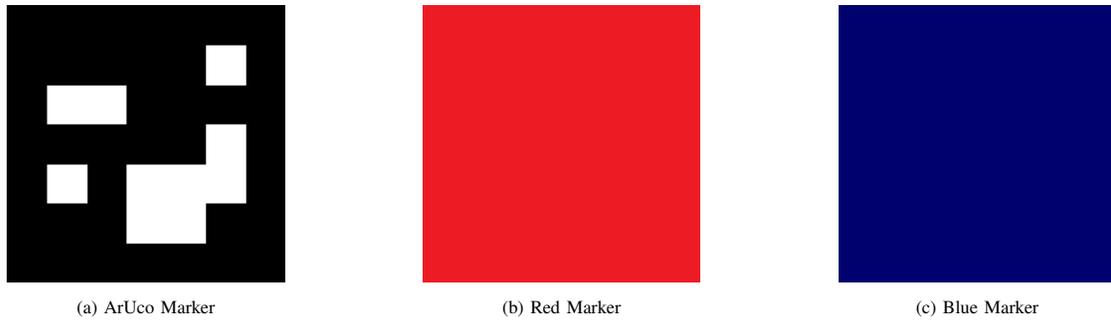


Fig. 1. Three Types of Markers.

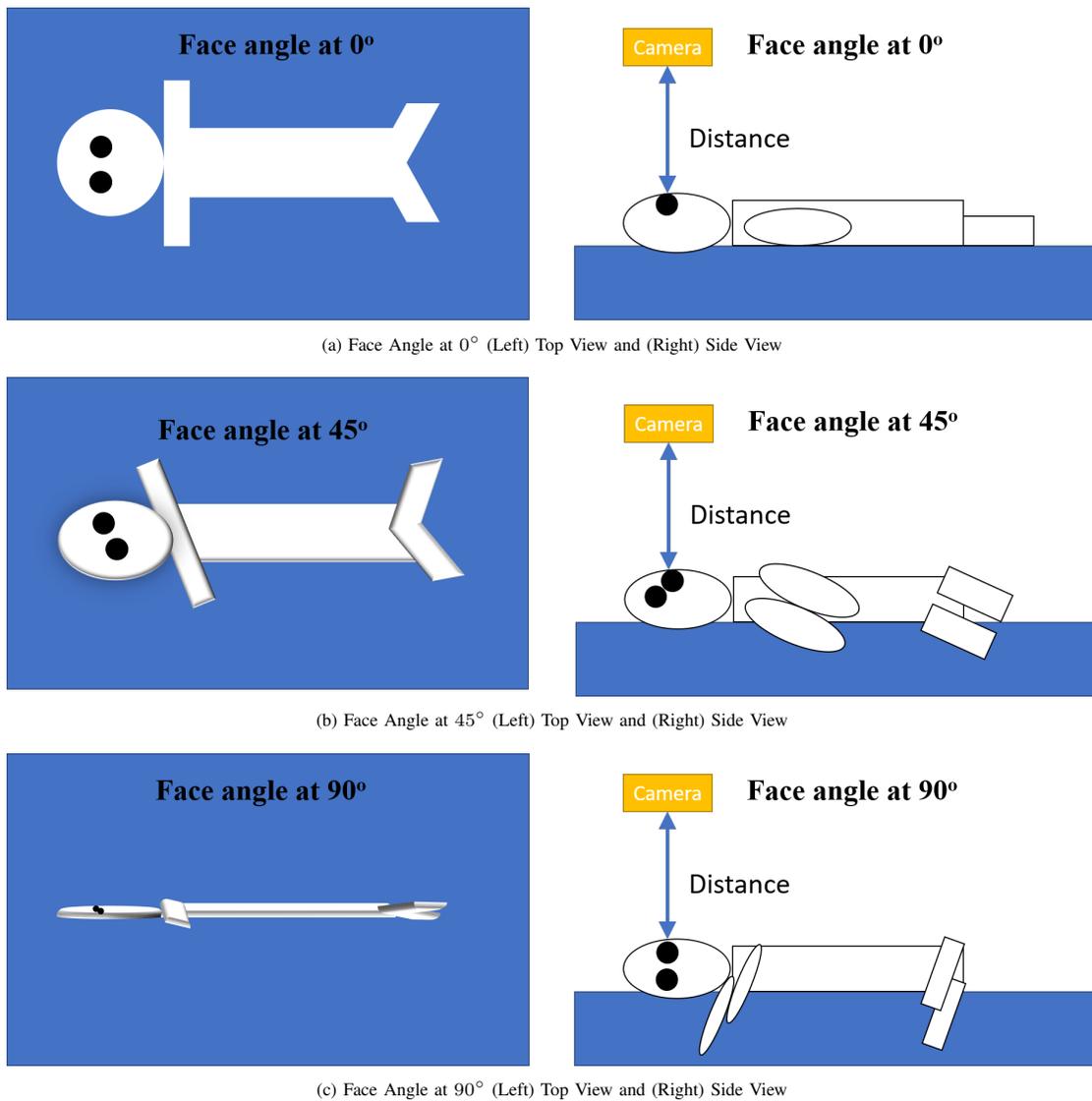


Fig. 2. Subject Poses at different Face Angles.

performed and the results were analyzed. The detection method using ArUco markers is found to be a good candidate. This method is robust to any practical disturbances on site and gives good reliability. Only drawback is the minimum marker size. It

is related to the image resolution. Hence, increasing the image resolution can solve this problem by changing the camera or reduce the subject to camera distance. A summary of pros and cons for each method is placed in Table III.

TABLE II. A PERFORMANCE SUMMARY FOR EACH DETECTION METHOD

Method	Performance in Percentage			
	Speed	Accuracy	Tolerance and Reliability	Resolution
a) ArUcoX2	100	70	11	78
b) ColorX2	100	0	0	0
c) FaceOnOff	100	33	33	100
d) FaceArUco	100	26	11	78
e) ColorArUco	100	44	3	58
f) FaceColor	100	20	6	61

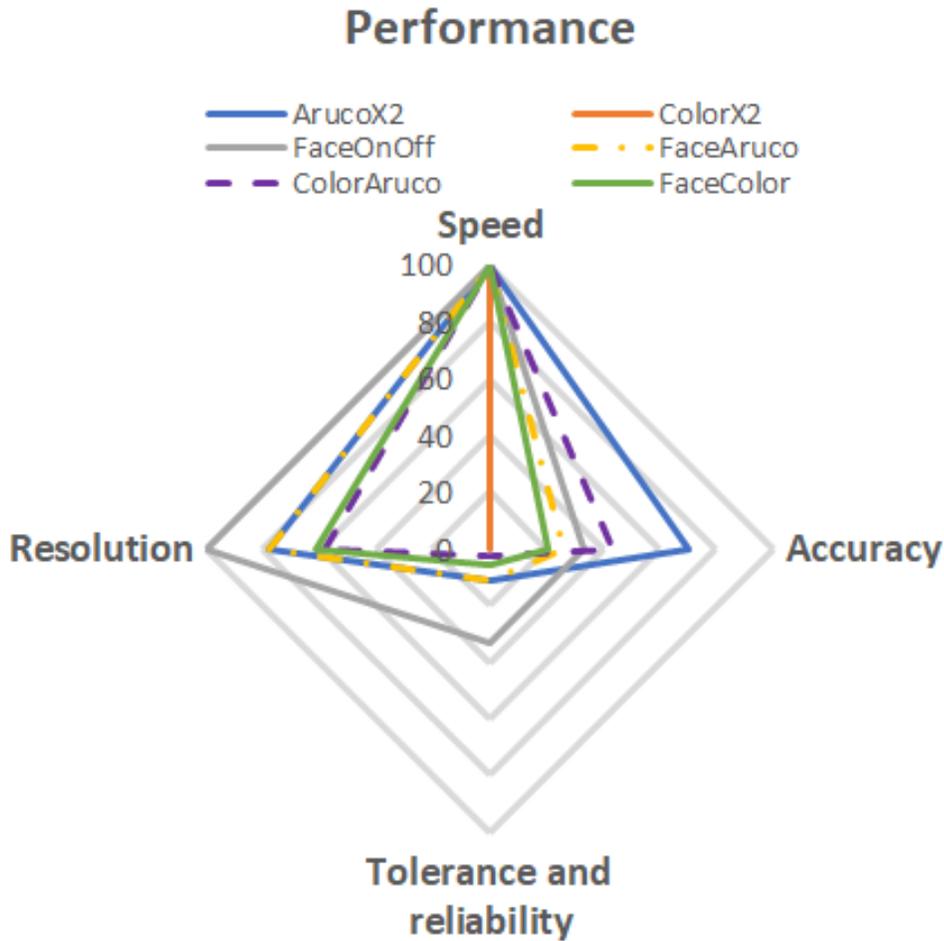


Fig. 3. Performance of Six Detection Methods.

TABLE III. A SUMMARY OF PROS AND CONS FOR EACH METHOD

Method	Pros	Cons
ArUco markers	Robust to any practical disturbances on site	Require good resolution image
Color markers	High detection speed	Sensitive to any practical disturbances on site
Face detection	Independent of image resolution	Critically require full face elements – two eyes, a nose and a mouth

Furthermore, the results also indicate that color markers are the worst method of detection. This is due to the acquired image does not have any color quality correction in order to have the fastest speed as much as possible. An improvement for this is to use image color correction tools to compensate

the light, saturation, tone and hue [41].

IV. CONCLUSION

To prevent a patient on the bed from trying to remove a nasogastric (NG) tube, an alarm system evaluates the patient's

gesture and warns the personal caretaker when the subject acts the risky pose by using visual technology. This system was evaluated a face detection and markers technique on Raspberry Pi 3 with a camera module. This work showed the detection and evaluate the risky pose of the subjects using visual technology, and also showed the comparison and evaluation of detection methods using visual technology. The experiment consists of five test conditions and six approaches, thereby there are 108 measurements times six methods, 648 measurements in total.

The results showed that a color marker method is the fastest method, however, it is very sensitive to marker size, subject to camera distance, light intensity, perspective distortion, background noise, and subject's face angle. A face detection method has the highest resolution regardless of the image resolution and the highest tolerance to the light intensity and background noise. On the contrary, this method requires the full subject's face including two eyes, one nose, and one mouth. An ArUco marker method is very robust to the light intensity, perspective distortion, background noise, and subject's face angle which give the highest accuracy. In contrast, it requires a good resolution image. As a result, The detection method using ArUco marker is found to be a good candidate. This method is robust to any practical disturbances on-site and gives good reliability. The only drawback is the minimum marker size. It is related to image resolution. Therefore, the ArUco marker is appropriate to be used in an alarm system preventing nasogastric (NG) tube removal of a patient.

V. LIMITATIONS AND FUTURE WORK

For the further work, a clinical study is needed to collect data and feedback for the patients with various conditions. In order to implement this work in the clinical study, the correlation between Aruco marker size and the subject to camera distance has to be evaluated. This is the most important factor to determine the warning. These factors can be compensated with high resolution camera. Another crucial issue is the performance under low light and dark conditions. Infrared light source and detectors can handle this issue.

ACKNOWLEDGMENT

This research has been supported by College Of Computing, Prince of Songkla University (Grant No. COC6504053S).

REFERENCES

- [1] J. Graham, N. Barnes, and A. S. Rubenstein, "The nasogastric tube as a cause of esophagitis and stricture," *The American Journal of Surgery*, vol. 98, no. 1, pp. 116–119, Jul. 1959.
- [2] T.-G. Wang, M.-C. Wu, Y.-C. Chang, T.-Y. Hsiao, and I. N. Lien, "The Effect of Nasogastric Tubes on Swallowing Function in Persons With Dysphagia Following Stroke," *Archives of Physical Medicine and Rehabilitation*, vol. 87, no. 9, pp. 1270–1273, Sep. 2006.
- [3] R. A. Sofferan, C. E. Haisch, J. A. Kirchner, and N. J. Hardin, "The nasogastric tube syndrome," *The Laryngoscope*, vol. 100, no. 9, pp. 962–968, 1990.
- [4] S. M, G. A, R. A, P. U, and K. A, "Spontaneous "Lariat loop" knotting of a nasogastric tube: prevention and management," *Tropical Gastroenterology*, vol. 35, no. 2, pp. 131–132, May 2015.
- [5] M. H. Trujillo, C. F. Fragachan, F. Tortoledo, and F. Ceballos, "Lariat Loop" Knotting of a Nasogastric Tube: An Ounce of Prevention," *American Journal of Critical Care*, vol. 15, no. 4, pp. 413–414, Jul. 2006.
- [6] M. I. Carrión, D. Ayuso, M. Marcos, M. P. Robles, M. A. de la Cal, I. Alía, and A. Esteban, "Accidental removal of endotracheal and nasogastric tubes and intravascular catheters," *Critical Care Medicine*, vol. 28, no. 1, pp. 63–66, Jan. 2000.
- [7] C. B. Pearce and H. D. Duncan, "Enteral feeding. Nasogastric, nasojejunal, percutaneous endoscopic gastrostomy, or jejunostomy: its indications and limitations," *Postgraduate Medical Journal*, vol. 78, no. 918, pp. 198–204, Apr. 2002, publisher: The Fellowship of Postgraduate Medicine Section: Review.
- [8] C. Best, "Caring for the patient with a nasogastric tube," *Nursing Standard*, vol. 20, no. 3, pp. 59–67, Sep. 2005.
- [9] J. B. Pillai, A. Vegas, and S. Brister, "Thoracic complications of nasogastric tube: review of safe practice," *Interactive CardioVascular and Thoracic Surgery*, vol. 4, no. 5, pp. 429–433, Oct. 2005.
- [10] S. Burns, M. Martin, V. Robbins, T. Friday, M. Coffindaffer, S. Burns, and J. Burns, "Comparison of nasogastric tube securing methods and tube types in medical intensive care patients," *American Journal of Critical Care*, vol. 4, no. 3, pp. 198–203, 05 1995. [Online]. Available: <https://doi.org/10.4037/ajcc1995.4.3.198>
- [11] Y. Kee, W. Brooks, R. Dhama, and A. Bhalla, "Evaluating the use of hand control mittens in post stroke patients who do not tolerate nasogastric feeding," *Cerebrovas Dis*, vol. 23, no. suppl 2, p. 93, 2007.
- [12] M. J. Popovich, J. D. Lockrem, and J. B. Zivot, "Nasal bridle revisited: an improvement in the technique to prevent unintentional removal of small-bore nasoenteric feeding tubes," *Critical care medicine*, vol. 24, no. 3, pp. 429–431, 1996.
- [13] D. Horsburgh, A. Rowat, C. Mahoney, and M. Dennis, "A necessary evil? interventions to prevent nasogastric tube-tugging after stroke," *British Journal of Neuroscience Nursing*, vol. 4, no. 5, pp. 230–234, 2008.
- [14] M. Dennis, "Nutrition after stroke," *British medical bulletin*, vol. 56, no. 2, pp. 466–475, 2000.
- [15] M. Dennis, S. Lewis, and C. Warlow, "Food trial collaboration routine oral nutritional supplementation for stroke patients in hospital (food): A multicentre randomised controlled trial," *Lancet*, vol. 365, no. 9461, pp. 755–763, 2005.
- [16] G. Kim, S. Baek, H.-w. Park, E. K. Kang, and G. Lee, "Effect of nasogastric tube on aspiration risk: results from 147 patients with dysphagia and literature review," *Dysphagia*, vol. 33, no. 6, pp. 731–738, 2018.
- [17] M. I. Carrión, D. Ayuso, M. Marcos, M. P. Robles, A. Miguel, I. Alía, and A. Esteban, "Accidental removal of endotracheal and nasogastric tubes and intravascular catheters," *Critical care medicine*, vol. 28, no. 1, pp. 63–66, 2000.
- [18] A. Brugnolli, E. Ambrosi, F. Canzan, L. Saiani, and N.-g. T. Group, "Securing of naso-gastric tubes in adult patients: a review," *International Journal of Nursing Studies*, vol. 51, no. 6, pp. 943–950, 2014.
- [19] C. W. Seder, W. Stockdale, L. Hale, and R. J. Janczyk, "Nasal bridling decreases feeding tube dislodgment and may increase caloric intake in the surgical intensive care unit: a randomized, controlled trial," *Critical care medicine*, vol. 38, no. 3, pp. 797–801, 2010.
- [20] A. Davoudi, K. R. Malhotra, B. Shickel, S. Siegel, S. Williams, M. Rupert, E. Bihorac, T. Ozrazgat-Baslanti, P. J. Tighe, A. Bihorac *et al.*, "Intelligent ICU for autonomous patient monitoring using pervasive sensing and deep learning," *Scientific reports*, vol. 9, no. 1, pp. 1–13, 2019.
- [21] E. Auvinet, F. Multon, A. Saint-Arnaud, J. Rousseau, and J. Meunier, "Fall detection with multiple cameras: An occlusion-resistant method based on 3-d silhouette vertical distribution," *IEEE transactions on information technology in biomedicine*, vol. 15, no. 2, pp. 290–300, 2010.
- [22] B. Mirmahboub, S. Samavi, N. Karimi, and S. Shirani, "Automatic monocular system for human fall detection based on variations in silhouette area," *IEEE transactions on biomedical engineering*, vol. 60, no. 2, pp. 427–436, 2012.
- [23] G. Mastorakis and D. Makris, "Fall detection system using Kinect's infrared sensor," *Journal of Real-Time Image Processing*, vol. 9, no. 4, pp. 635–646, 2014.

- [24] M. Humenberger, S. Schraml, C. Sulzbachner, A. N. Belbachir, A. Srp, and F. Vajda, "Embedded fall detection with a neural network and bio-inspired stereo vision," in *2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*. IEEE, 2012, pp. 60–67.
- [25] C. Rougier, J. Meunier, A. St-Arnaud, and J. Rousseau, "3d head tracking for fall detection using a single calibrated camera," *Image and Vision Computing*, vol. 31, no. 3, pp. 246–254, 2013.
- [26] K. Makantasis, E. Protopapadakis, A. Doulamis, L. Grammatikopoulos, and C. Stentoumis, "Monocular camera fall detection system exploiting 3d measures: a semi-supervised learning approach," in *European Conference on Computer Vision*. Springer, 2012, pp. 81–90.
- [27] Y. Song, D. Demirdjian, and R. Davis, "Tracking body and hands for gesture recognition: NATOPS aircraft handling signals database," in *2011 IEEE International Conference on Automatic Face Gesture Recognition (FG)*, Mar. 2011, pp. 500–506.
- [28] N. Mohsin, S. Payandeh, D. Ho, and J. P. Gelinis, "Bluetooth Low Energy Based Activity Tracking of Patient," in *2018 15th International Conference on Control, Automation, Robotics and Vision (ICARCV)*, Nov. 2018, pp. 1991–1996.
- [29] D.-S. Kim, J. Kim, S.-H. Kim, and S. K. Yoo, "Design of RFID based the Patient Management and Tracking System in hospital," in *2008 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, Aug. 2008, pp. 1459–1461, ISSN: 1558-4615.
- [30] R. Spin-Neto, L. H. Matzen, L. Schropp, E. Gotfredsen, and A. Wenzel, "Detection of patient movement during CBCT examination using video observation compared with an accelerometer-gyroscope tracking system," *Dentomaxillofacial Radiology*, vol. 46, no. 2, p. 20160289, Feb. 2017.
- [31] R. Mardiyanto, M. F. R. Utomo, D. Purwanto, and H. Suryoatmojo, "Development of hand gesture recognition sensor based on accelerometer and gyroscope for controlling arm of underwater remotely operated robot," in *2017 International Seminar on Intelligent Technology and Its Applications (ISITIA)*, Aug. 2017, pp. 329–333.
- [32] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, vol. 1, Dec. 2001, pp. 1–I, ISSN: 1063-6919.
- [33] P. Viola and M. J. Jones, "Robust Real-Time Face Detection," *International Journal of Computer Vision*, vol. 57, no. 2, pp. 137–154, May 2004. [Online]. Available: <https://doi.org/10.1023/B:VISI.0000013087.49260.fb>
- [34] S. Liao, X. Zhu, Z. Lei, L. Zhang, and S. Z. Li, "Learning Multi-scale Block Local Binary Patterns for Face Recognition," in *Advances in Biometrics*, ser. Lecture Notes in Computer Science, S.-W. Lee and S. Z. Li, Eds. Berlin, Heidelberg: Springer, 2007, pp. 828–837.
- [35] S. Garrido-Jurado, R. Muñoz-Salinas, F. J. Madrid-Cuevas, and M. J. Marín-Jiménez, "Automatic generation and detection of highly reliable fiducial markers under occlusion," *Pattern Recognition*, vol. 47, no. 6, pp. 2280–2292, Jun. 2014.
- [36] M. Kalaitzakis, B. Cain, S. Carroll, A. Ambrosi, C. Whitehead, and N. Vitzilaios, "Fiducial Markers for Pose Estimation," *Journal of Intelligent & Robotic Systems*, vol. 101, no. 4, p. 71, Mar. 2021.
- [37] S. Garrido-Jurado, R. Muñoz-Salinas, F. J. Madrid-Cuevas, and R. Medina-Carnicer, "Generation of fiducial marker dictionaries using Mixed Integer Linear Programming," *Pattern Recognition*, vol. 51, pp. 481–491, Mar. 2016.
- [38] F. J. Romero-Ramirez, R. Muñoz-Salinas, and R. Medina-Carnicer, "Speeded up detection of squared fiducial markers," *Image and Vision Computing*, vol. 76, pp. 38–47, Aug. 2018.
- [39] B. Woo, Y. Uh, K. Lim, Y. Choi, and H. Byun, "Illumination invariant color segmentation method based on cluster center tree for traffic sign detection," in *Proceedings of the 9th International Conference on Ubiquitous Information Management and Communication*, ser. IMCOM '15. New York, NY, USA: Association for Computing Machinery, Jan. 2015, pp. 1–5.
- [40] P. Wonghabut, J. Kumphong, R. Ung-arunyawee, W. Leelapatra, and T. Satiennam, "Traffic Light Color Identification for Automatic Traffic Light Violation Detection System," in *2018 International Conference on Engineering, Applied Sciences, and Technology (ICEAST)*, Jul. 2018, pp. 1–4.
- [41] B. Muhammad and S. A. Rahman Abu-Bakar, "A hybrid skin color detection using HSV and YCbCr color space for face detection," in *2015 IEEE International Conference on Signal and Image Processing Applications (ICSIPA)*, Oct. 2015, pp. 95–98.

Multileveled ALPR using Block-Binary-Pixel-Sum Descriptor and Linear SVC

B. Lavanya¹, G. Lalitha²
Associate Professor¹, Research Scholar²
Department of Computer Science
University of Madras
Chennai-25, Tamil Nadu, India

Abstract—Automatic license plate recognition (ALPR) is essential component of security and surveillance. ALPR mainly aims to detect and prevent the crime and fraud activities; it also plays an important role in traffic monitoring. An algorithm is proposed for recognizing license plate candidates. The proposed work aimed to recognize the license plate of a car. Proposed work is designed in multilevel for more accurate License Plate (LP) recognition, At level 1 algorithm produced 93.5% accuracy and in level 3 algorithm gives 96% accuracy. For training and testing purpose, LP images were used from Medialab cars dataset, kaggle car dataset and goggle map images. These images in the dataset is formulated at various angles and illumination. Proposed algorithm for LP recognition is done by using the Block Binary Pixel descriptors (BBPS) and Linear Support Vector Classification (SVC). Proposed algorithm is novel and produces higher accuracy in minimal processing time of an average 0.42 milliseconds with 96% accuracy when compared with state-of-the art methods.

Keywords—BBPS – Block Binary Pixel Sum; ALPR - Automatic License Plate Recognition; ROI - Region of Interest; SVC - Support Vector Classification; LP - License Plate

I. INTRODUCTION

ALPR is a active, popular and interesting topic in image processing [1], development in license plate recognition have received much attention towards English license plate recognition [2]. It also plays an important role in intelligent transport system [3] [4] [5], border control, toll collection, traffic management [6] [7], it is used in information and communication technology [8]. Computer vision plays an major part in extracting ROI from an image [9]; license plate recognition is done by extracting number plate from images [3] where the Region of Interest(ROI) is license plate of an vehicle, the main task is to extract characters from license plate [10], each vehicle number plate carries different font style and size [11].Surveillance camera is fixed everywhere, earlier it is fixed in important places like malls, hospitals and in sensitive places, but now it's fixed everywhere even people prefer to fix surveillance- camera in their places for their safety. Since many algorithms exists for ALPR but still it lags the accuracy due to various factors like images were captured at night time, tilted and blurred images, so there is a need for an efficient algorithm to improve accuracy. Existing algorithms were proposed using various pre-processing steps followed by feature extraction, machine learning and deep learning techniques. Before extracting features from image pre-processing techniques should be applied on image. Preprocessing the input image which includes conversion of input image to a

gray scale image and pre-processing is applied to enhance the details present in the image, in order to highlight the ROI and prune the non-ROI. Then morphological operations like erosion, dilation, tophat, blackhat are applied to enrich input image feature, which supports system proficiency [12]. For recognizing character from image the popular method called optical character recognition with tesseract, an emerging concept is used. ALPR system follows the sequence of steps

- Acquisition of input image
- Localizing license plate
- Segmenting and slicing license plate characters
- Recognizing license plate characters

Since many algorithm exists for ALPR, but it still lags to recognize license plate accurately. So there is a need for an better algorithm to recognize license plate of vehicle accurately even the image captured in uneven lightning, blurred [13]. Proposed work overcomes all those existing disadvantages by recognizing license plate accurately since training of license plate recognition is done by extracting real time license plate images, which tends for more accurate precision. Proposed work is done in multilevel so as to improve the accuracy of proposed algorithm. Many existing algorithm exists for ALPR but its still have many unsolved queries like, it lags to work on images captured at night and dark environment, to recognize license plate of images taken from far distance, to work on skewed and blurred images. It also produces false positive, false negative results for many license plates and existing works fails to predict between characters of similar strokes. So has to solve existing algorithm flaws proposed algorithm is designed in a novel way to overcome all those existing flaws and to recognize license plate accurately from complex backgrounds.

II. RELATED WORK

ALPR plays an important part in border scrutiny, confirming safeguards, and dealing with vehicle-related crime. ALPR system makes use of deep learning techniques using a convolutional neural network (CNN). The CNN assigns significance to numerous features of the image and differentiating them from each other. According to researchers point of view CNN works good for LP character recognition [3]. ALPR received much consideration for the English license plate. ALPR addresses more frightening traffic dealing with scanty road-safety processes. Architecture of ALPR for predicting vehicle LP regions

prior to Vehicle License Plate (VLP) in the ideal is proposed to eliminate false positives causing in higher prediction accuracy of VLP [2]. ALPR plays essential part in current transport organizations such as traffic monitoring and vehicle violation detection. In real-world scenarios, LP recognition faces many contests and is diminished by unrecognized interfering such as weather or lighting conditions. Machine learning based ALPR solutions has been wished-for to resolve such tests in recent years. However, most of the algorithm does not yield convincing results since their consequences are appraised on small or simple datasets that lack varied surrounds, or it require potent hardware to accomplish a practical frames-per-second in real-world solicitations [14]. In ALPR initially, license plate location will be determined. Then, in the second phase, enhancement can be done by applying Gaussian function for filtering. Next, edges in the image located so that LP location can be spotted. Then tilting and plate rotation and affine transformation are applied if the input image was captured in tilted and slanting position or angles. Then by neural network concepts LP characters extracted from LP images [15]. ALPR [16] system plays vital role in Intelligent Transport System (ITS) their major goal was traffic controlling. ALPR for vehicles is a main part of ITS. ALPR send images to a server for LP recognition. To decrease stays and bandwidth usage during images communication, an Edge-AI-based Real-time (ER) ALPR ER-ALPR was proposed, in which an AGX XAVIER entrenched system is embedded on the edge of a camera to attain real-time image input to an AGX edge device and to enable real-time automatic LP candidate recognition. To measure LP characters and styles in a precise setting, the ER-ALPR scheme smears the following methods: (1) image pre-processing (2) You Only Look Once v4-Tiny (YOLOv4-Tiny) [17] for LP prediction; (3) virtual judgment line for determining whether a license plate frame has passed; (4) the proposed modified YOLOv4 (M-YOLOv4) for LP candidate recognition; and (5) a logic ancillary ruling scheme for refining LP recognition. ER-ALPR system was complete in real-life test surroundings in Taiwan. Many state-of-art methodologies exists for recognizing text from various environment, but none of the algorithm recognizes accurately, it lags to differentiate between similar characters, existing algorithm lags to recognize license plate characters from complex environments.

III. THE PROPOSED WORK

The proposed algorithm consists of three levels and each level has different phases

Level 1: Training phase

Input: Load various input font images.

- Initialize the BBPS Descriptor
- Train separately both character and digit classifier using Linear SVC to recognize LP accurately

Output: Dump the character and digit classifier as a file (pickle file).

Testing phase

Input: Load the pickle file (both character and digit classifier) and dataset images.

- Initialize the BBPS Descriptor

Output: Print the license plate text on the image

Level 2: gathering labeled data

This phase is used to train the character and digit classifier for more accurate prediction of LP characters.

Input: Localized license plate images

- Loop over the characters present in each localized LP Images
- Grab each character, create a directory and store it for advanced training

Output: Independent folder created for each character and digits; which is used for advanced training for LP recognition.

Level 3: Advance training phase

Input: Load the sample output generated in level 2 (various segmented image stored in labeled folders)

- Repeat level 1 training phase

Output: Dump the character and digit classifier as a file (advance pickle File). Testing phase

Input: Load the advance pickle file (both character and digit classifier) and dataset images and initialize the BBPS descriptor.

Output: Print the license plate text on the image

A. The Proposed Work (Level 1)

Aim of the proposed work is to propose a powerful algorithm to recognize license plate candidates accurately. Here BBPS descriptor is used to propose an algorithm to extract features of license plate characters. Initially various font images were downloaded from web which is used to train the classifier. Linear SVC is used to train the digit and character classifier separately to obtain higher accuracy. Training dataset were constructed by segmenting LP characters from dataset Images and stored in various labelled folders, which is used in testing phase.

1) *Methodology*: The Objective of the proposed work is to recognize LP images. So a different font style characters and digit dataset images downloaded from web which look alike Fig 1 and used for training the algorithm. For training eight different font style applied image collected from web. Each font displays the characters A-Z and the numbers 0-9. Since we have eight images, we have eight examples for each letter and digit. Proposed algorithm extracts BBPS features from the characters and trained using linear SVC : one for letter recognition and a separate one for digit recognition.



ABCDEFGHIJKLMN OPQRSTUVWXYZ0123456789
ABCDEFGHIJKLMN OPQRS TUVWXYZ 0123456789
ig. 1. Training set for License Plate Images

2) *Block-Binary-Pixel-Sum Descriptor(BBPS)*: BBPS descriptor [18], is used for license plate character recognition. BBPS functions by separating an image into non-overlapping $M \times N$ pixel blocks, by applying BBPS descriptor on image ; image will be subdivided on three basis one is 3×3 regions, another one is 2×3 regions and last one is 3×2 sub division as shown in Fig 2. In BBPS target size of the image is fixed, canonical size of ROI is resized so that all images can maintain consistent representation and quantification of each character from dataset images. Input image is converted to a binary image, with pixels corresponding to the character having an intensity > 0 , and pixels corresponding to the background set to 0.

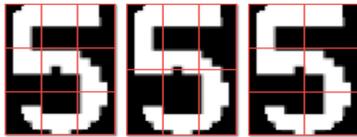


Fig. 2. Example for BBPS Descriptor

3) *Training Classifiers*: License plate has two groups of characters letters followed by digits. In order to obtain better accuracy, both the digit and character classifier should be trained separately. Here linear SVC is used to train a classifier for classification or prediction. Character contains alphabets from string a-z and the digits 0-9. In testing phase all font images will be loaded for training the classifier, convert it to a grayscale image, and threshold operator applied on image so the image look like Fig. 3.



Fig. 3. Training Image after Applying threshold Operator

After thresholding image, contour detection was applied on image to sort our contours from left-to-right, bounding box computed for each of the sorted contours. For each of the sorted contours ROI extracted and passed to BBPS i.e., LP. Training dataset image has characters followed by digits so if value < 26 means it is character otherwise digit. Respective digit and characters updated in appropriate labeled lists. Each character and digit is classified using linear SVC, it will fit the data which is loaded into it will provide, a finest fitting hyperactive plane that rifts or classifies the statistics which feed in to it, linear SVC will dump the data along with its label value as a model. Linear SVC created model will be dumped into pickle file, at the end of training phase two separate pickle file will be created one for character and another for digit. Pickle file which is used for serializing and de-serializing an object structure, which maintains program state across sessions, or transport data over the network. Here we dumped both the classifier in pickle file in order to transmit training data to testing phase. Fig 5 depicts the work flow of level 1 training phase.

4) *Testing Phase*: In this phase two pickle file created in training phase will be loaded along with the testing dataset. From the localized license plate images [12], images were looped over the LP regions bounding box, for each of extracted characters from LP regions [19] applied BBPS descriptor, LP images always carry character followed by digits, So first character classifier is applied to predict features of LP characters followed by digit classifier, length of the bounding box characters were checked if length of the character greater than zero i.e., $\text{len}(\text{chars}) > 0$ and compute the center of LP regions so has to print the detected strings parallel to the LP region. Fig. 6 depicts the work flow of level 1 testing phase.



Fig. 4. License Plate Recognition via Proposed Algorithm

After applying the proposed algorithm, LP image look like Fig 4, proposed algorithm gives a good accuracy and prediction rate. Fig 12 shows the successfull LP recognition in level 1. In Fig 13 LP recognized falsely their character M is wrongly recognized has N since both characters looks similar our proposed classifier lags in differentiating nearest character. In Fig 13 digits were misclassification 1 is recognized has 9 so an enhanced proposed algorithm needed and for such wrong redictions. In order to overcome false prediction characters and to improve accuracy, training classifier were trained using real-time license plate images. License plate characters were extracted and labeled has a sample of license character examples from the license plate dataset, both digit and character classifier has been re-trained on top of the BBPS feature representations from the more “real-world” dataset, testing done with the re-trained classifiers which lead to obtain higher character identification accuracy.

B. The Proposed Work - Level 2

- In order to improvise accuracy training classifier done using real-time license plate images.
- LP characters; extracted and labeled has a sample of license character examples from our license plate dataset.
- Both digit and character classifier will be re-trained on top of the BBPS feature representations from the more “real-world” dataset.
- Testing done with the re-trained classifiers which lead to obtain higher character identification accuracy.

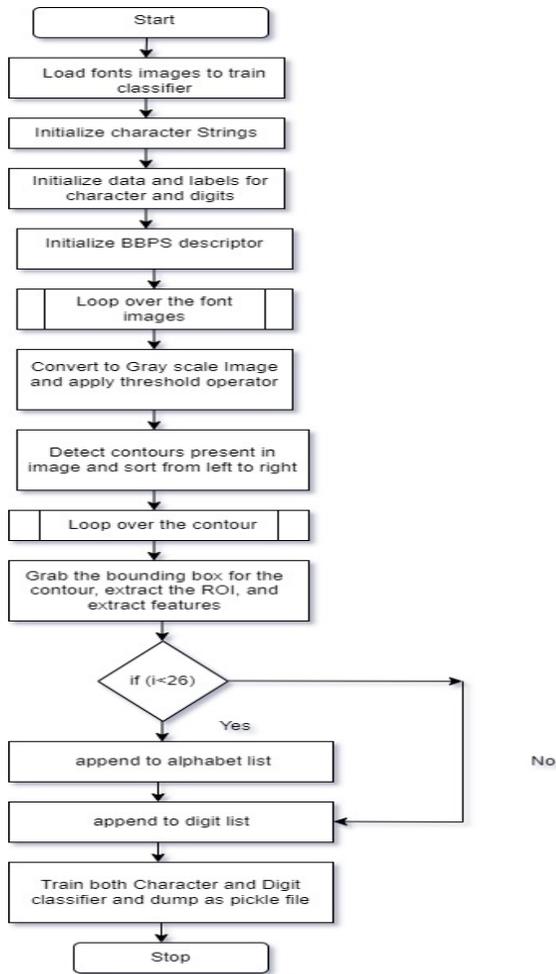


Fig. 5. Flow of Work of Level 1 Training Phase

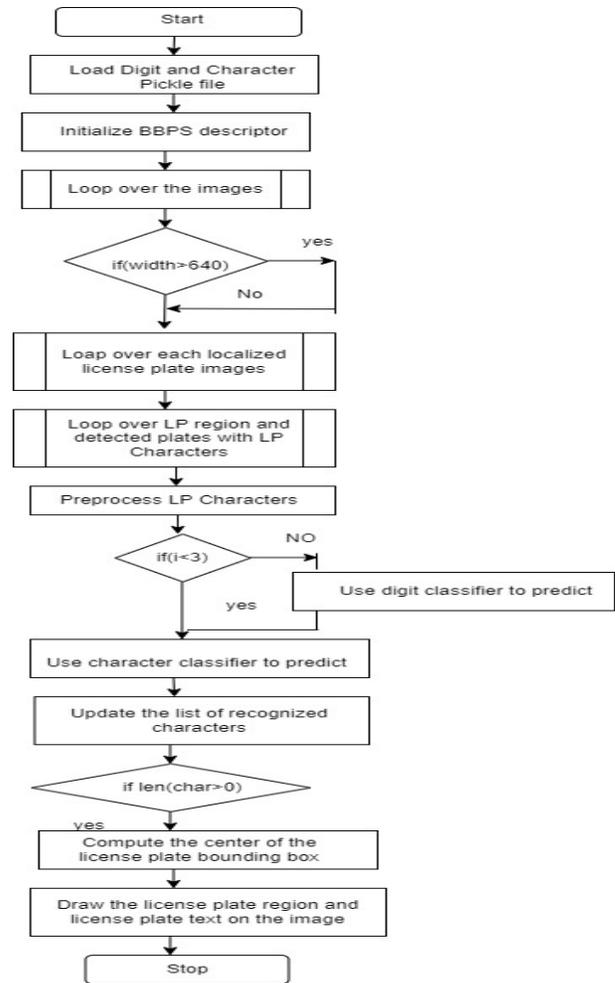


Fig. 6. Flow of Work of Level 1 Testing Phase

Aim for retraining classifier is to obtain a higher accuracy; here classifier is retrained using real-world character examples using LP dataset. LP dataset which used to training is a real-world LP car image dataset so that it can recognize LP characters of real world vehicles. Fig. 7 depicts the work flow of level 2. From a localized bounding box generated images [12]. In second phase initially will randomly select 50% images for training and 50% images for testing. LP localized image with bounding box will be loaded [12], will loop over the LP region with bounding boxes and extract LP characters [19], in that each of the characters from LP images will be displayed and wait for key press event to manually label each of these LP characters like Fig 8. Once characters extracted from LP, it will wait for key press event. If the ‘/’ (backtick/tilde) key was pressed, proposed work will ignore the character which is built to ignore falsely detected characters or “noise” in the LP image. Other key press represents confirmation of the character without labelled data; meanwhile appropriate key pressed for characters on screen those characters will be stored with correct label like Fig. 9. Once the LP character has been labelled, it will be written in to disk using a directory structure like filename followed by extension like Fig. 10.

Once characters extracted from license plate, once backtick

or tilde key was pressed, proposed work will ignore the character which is built to ignore falsely detected characters or noise in the LP image. Other characters other than other key backtick or tilde key pressed which represents confirmation of the character without labelled data; meanwhile appropriate key pressed for characters on screen those characters will be stored with correct label like Fig. 8. Once the LP character has been labelled, it will be written in to disk using a directory structure like character name/example name.png

By pressing appropriate key while labelling data, character will be saved to folder as like Fig. 10. While comparing with our previous level contrived LP fonts samples these labelled LP characters are certainly representative of real-world LP characters like Fig. 11. In Level 2 labeled characters directory was created with real-time LP images, since web downloaded font image were trained and used in level 1 which lags to give exact recognition at some cases since there is lots of difference between web downloaded image and real LP images, so if training classifier also done with real LP images which gives good accuracy when testing with real LP images. In level 3 from LP dataset images using BBPS extract features, and retrain the character and digit classifier.

V. RESULT AND DISCUSSION

The proposed work license plate recognition is done using medialab dataset, kaggle car image dataset and Google map images. Proposed algorithm build using BBPS descriptor implemented in anaconda python. The experiment is done in laptop with intel core i5, 2.70 GHz, 8 GB RAM windows 10-64-bit OS. On execution time all the images were rescaled and maintained the size of 640 pixels with unchanged aspect ratio. Proposed algorithm gives very good accuracy compared to existing works. Proposed work is novel and works better than the state-of-the-art methodologies. Existing works lags to recognize license plate from the blurred, sweked images captured in uneven lightning and fails to recognize license plate characters of images taken from longer distance, proposed algorithm works on all the above cases so its a novel and works better than state-of-the-art methodologies. Proposed algorithm is simple to understand, easy to implement in real time environment which gets executed in 0.42 milliseconds. Proposed algorithm is proposed in such a way to overcome all the flaws of existing methodologies.

Fig. 12 depicts the successful true positive result in level 1. Here license plate were successfully recognized whereas in Fig. 13 in left side image all candidates were recognized correctly except the second character candidate which is recognized has N instead of M. In Fig. 13 all character candidates were recognized correctly except a digit '1' which is recognized has '9'. All these misclassification occur because all these candidates share similar features.



Fig. 12. Successful License Plate Recognition in Level 1



Fig. 13. False Prediction in Level 1

In level 1 training of LP characters were done using different font images downloaded from internet, level 1 yields good accuracy but it fails to recognize a few cases, in order to overcome such flaws, training of classifier is done by gathering characters from dataset images like Fig 14. In Fig. 15 depicts the wrongly predicted license plate images in level 1 which is correctly recognized in level 3.



Fig. 14. Gathering LP Characters in Level 2



Fig. 15. Wrongly Recognized Vehicle LP in Level 1 Predicted Correctly in Level 3

VI. PSEUDOCODE

Level 1: Training Phase

initialize characters string, data and labels for the alphabet and digits

initialize BBPS descriptor

loop dataset images, convert to grayscale

and apply thresholding

detect contours; sort from left to right

loop over contours, grab bounding box of contour

extract ROI

if $i < 26$

it is character

else

digit

train character and digit classifier

dump the character and digit classifier to a file

Testing phase

def testing phase()

load the character and digit classifiers created in training

phase

initialize BBPS descriptor

loop dataset images

```

if width > 640,
    resize(image)
detect the license plates [12] and characters [19]
loop over detected plates and characters
preprocess it
    if i > 3
        call digit classifier
    else
        call character classifier
    update the text of recognized characters
if len(chars)>0
    compute the center of LP bounding box
    draw the characters which are predicted

```

above bounding box

Level 2: gathering LP characters

```

randomly select a portion of the images
loop dataset images
    if width > 640
        resize(image)
    detect the license plate [12] and characters [19]
    draw bounding box around recognized
    license plate
    loop over the characters
    display the character
    if actual != segmented character
        print(ignore)
        continue
    construct the path to the output directory
    if output directory not exist, create directory
    write the labeled character to file

```

Level 3: Advanced training phase

```

initialize the BBPS descriptor
initialize the data and labels for the alphabet and digits
loop over the sample dataset created in level 2
    extract the images
    loop over the images
        load the character, convert it to grayscale
        preprocess it
        if the character is digit
            append digit data and label
        otherwise
            append alphabet data and label

```

train both advanced character and digit classifier

dump the character and digit classifier to file

Advanced testing phase

```

Input: Load advanced character and digit classifier
call testing phase()

```

VII. PERFORMANCE EVALUATION

In the proposed work in level 1 few LP character candidates were predicted wrongly, so to overcome such flaws training images were created using medialab, kaggle car dataset and Google map images. To evaluate the performance of proposed system accurately, here we have trained and tested the proposed system with three dataset images, which has images captured at various angles, various lightning condition. In a Tunisian dataset, [20] which is used for calculating accuracy for LP recognition, which comprised of true positive and true negative divided by true positive, true negative, false positive and false negative boundary boxes. Table 1 depicts different dataset images used, the performance metrics of proposed

system is calculated in terms of measuring precision, recall, f-measure. Proposed algorithm recognize dataset images more accurately than the existing algorithm, it recognizes images captured at challenging environment and angles. Performances of proposed algorithm is produced in table format. If the license plate correctly predicted from license plate images which come under True Positive (TP), if non-license plate region recognized has license plate which called as False Positive (FP). Since the proposed work is done in multilevel performance evaluation has been done for each level independently in Tables 2, 3 and 4 depicts the true positive, false positive values from all the three levels of proposed work.

$$Precision = \frac{TP}{TP + FP} \quad (1)$$

$$Recall = \frac{TP}{TP + FN} \quad (2)$$

$$F - measure = \frac{2 * Precision * Recall}{Precision + Recall} \quad (3)$$

TABLE I. DIFFERENT DATASET IMAGES

Dataset Name	No.of.Images
Medialab	139
Kaggle dataset	55
Google map images	20

TABLE II. CONFUSION MATRIX FOR LEVEL 1

	Predicted Negative (PN)	Predicted Positive (PP)
Actual Positive(AP)	11	17
Actual Negative(AN)	0	204

TABLE III. CONFUSION MATRIX FOR LEVEL 2

	Predicted Negative (PN)	Predicted Positive (PP)
Actual Positive(AP)	0	7
Actual Negative(AN)	0	207

TABLE IV. CONFUSION MATRIX FOR LEVEL 3

	Predicted Negative (PN)	Predicted Positive (PP)
Actual Positive(AP)	3	13
Actual Negative(AN)	0	198

In prediction phase if both the license plate character candidate and predicted character candidate were same means its true positive, if non LP region predicted has LP candidate means its false positive, if non license plate region not recognized has license plate region its true negative, if non license

plate region detected has license plate region means false negative. Equations for precision, recall, f-measure given in eq(1),(2),(3). Precision, recall, f-measure were metrics which is used to test performance of an algorithm. Precision is a measure which can be used when the count of false positive is less. Recall is a metrics which is used to improve the prediction rate when it lag to produce satisfying result in precision phase; this metrics can be used when the value of false negative is less. F1 measure can be used when the accuracy seek the balance between precision and recall, when class distribution is uneven i.e more actual negatives. Table 5 depicts the precision, recall, f-measure score of proposed work. Table 6 depicts the proposed methodology accuracy with existing works.

TABLE V. PRECISION, RECALL, F-MEASURE VALUES

	Precision	Recall	F1-score
Level-1	92.3	94.8	93.5
Level-2	96.7	100	98.3
Level-3	93.8	98.5	96

TABLE VI. COMPARISON WITH EXISTING WORKS

S.no	Author's	Precision	Recall	F-measure
1	Proposed algorithm	93.8%	98.5%	96%
2	Omar et al [10].,	94.43%	92.10%	91.01%
3	Al-Shemarry et al [21].,	91.6%	87.1%	89.33%

Proposed work evaluated with popular dataset like medi-alab, kaggle, and Google map images. All images were used for both training the classifier and testing license plates. Fig. 16 depicts the final outcome by kaggle dataset images and Fig. 17 is a sample output for successful license plate recognition for google map images.



Fig. 16. Successful License Plate Recognition using kaggle Dataset

VIII. CONCLUSION

ALPR is an interesting hot topic, many algorithm exists for ALPR but it still lags on recognizing license plate characters from complex environment, by overcoming all the backdrops in existing system a novel and efficient algorithm produced here to recognize license plate characters accurately from complex environment. In this proposed work, license plate is recognized from media-lab,kaggle car dataset images and



Fig. 17. Successful License Plate Recognition using Goggle Map Images

from Google map images. Many existing algorithm exist for LP recognition but it lags to provide good accuracy due to various factors like images were captured in uneven lightning, taken at tilted position and blurred images. Existing algorithms are tedious, time consuming and it requires labeled data, those were expensive to compute, and highly time consuming. Proposed algorithm is novel, built using BBPS descriptor and linear SVC which is super-fast, run easily in real-time, and its simple to comprehend, its inexpensive to compute, and executes in 0.42 milliseconds. In level 1 proposed work produced 93.5% accuracy, In level 2 gathered training dataset images and which gave 98.3% in recognizing and training dataset images which is used as a input to level 3. level 3 produced an enhanced better results than proposed level 1 and other existing methodologies. As a future work license plate of regional language images will be recognized.

ACKNOWLEDGMENT

This research did not receive any specific grant from funding agencies in the public,commercial,or not-for-profit sectors.

REFERENCES

- [1] Z. Selmi, M. B. Halima, U. Pal, and M. A. Alimi, "Delp-dar system for license plate detection and recognition," *Pattern Recognition Letters*, vol. 129, pp. 213–223, 2020.
- [2] M. Onim, S. Hassan, H. Nyeem, K. Roy, M. Hasan, A. Ishmam, M. Akif, A. Hoque, and T. B. Ovi, "Blpnet: A new dnn model and bengali ocr engine for automatic license plate recognition," *arXiv preprint arXiv:2202.12250*, 2022.
- [3] P. Kaur, Y. Kumar, S. Ahmed, A. Alhumam, R. Singla, and M. F. Ijaz, "Automatic license plate recognition system for vehicles using a cnn," *CMC-COMPUTERS MATERIALS & CONTINUA*, vol. 71, no. 1, pp. 35–50, 2022.
- [4] J. Liang, G. Chen, Y. Wang, and H. Qin, "Egsanet: edge-guided sparse attention network for improving license plate detection in the wild," *Applied Intelligence*, vol. 52, no. 4, pp. 4458–4472, 2022.
- [5] W. Al Faqheri and S. Mashohor, "A real-time malaysian automatic license plate recognition (m-alpr) using hybrid fuzzy," *International Journal of Computer Science and Network Security*, vol. 9, no. 2, pp. 333–340, 2009.
- [6] R. Laroca, L. A. Zanlorensi, G. R. Gonçalves, E. Todt, W. R. Schwartz, and D. Menotti, "An efficient and layout-independent automatic license plate recognition system based on the yolo detector," *arXiv preprint arXiv:1909.01754*, 2019.

- [7] K. Kluwak, J. Segen, M. Kulbacki, A. Drabik, and K. Wojciechowski, "Alpr-extension to traditional plate recognition methods," in *Asian Conference on Intelligent Information and Database Systems*. Springer, 2016, pp. 755–764.
- [8] P. Agarwal, K. Chopra, M. Kashif, and V. Kumari, "Implementing alpr for detection of traffic violations: a step towards sustainability," *Procedia Computer Science*, vol. 132, pp. 738–743, 2018.
- [9] A. Ashrafee, A. M. Khan, M. S. Irbaz, A. Nasim, and M. Abdullah, "Real-time bangla license plate recognition system for low resource video-based applications," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2022, pp. 479–488.
- [10] S. Du, M. Ibrahim, M. Shehata, and W. Badawy, "Automatic license plate recognition (alpr): A state-of-the-art review," *IEEE Transactions on circuits and systems for video technology*, vol. 23, no. 2, pp. 311–325, 2012.
- [11] W. Al Faqheri and S. Mashohor, "A real-time malaysian automatic license plate recognition (m-alpr) using hybrid fuzzy," *International Journal of Computer Science and Network Security*, vol. 9, no. 2, pp. 333–340, 2009.
- [12] G. Lalitha and S. Gopinathan, "An effective algorithm for detecting and localizing license plate images," *Journal of Fundamental & Comparative Research (Shodhsamhita)*, 2021.
- [13] R. M. Khoshki and S. Ganesan, "Improved automatic license plate recognition (alpr) system based on single pass connected component labeling (ccl) and reign property function," in *2015 IEEE International Conference on Electro/Information Technology (EIT)*. IEEE, 2015, pp. 426–431.
- [14] S.-R. Wang, H.-Y. Shih, Z.-Y. Shen, and W.-K. Tai, "End-to-end high accuracy license plate recognition based on depthwise separable convolution networks," *arXiv preprint arXiv:2202.10277*, 2022.
- [15] O. Akbarzadeh, M. R. Khosravi, and L. T. Alex, "Design and matlab simulation of persian license plate recognition using neural network and image filtering for intelligent transportation systems," *ASP Transactions on Pattern Recognition and Intelligent Systems*, vol. 2, no. 1, pp. 1–14, 2022.
- [16] C.-J. Lin, C.-C. Chuang, and H.-Y. Lin, "Edge-ai-based real-time automated license plate recognition system," *Applied Sciences*, vol. 12, no. 3, p. 1445, 2022.
- [17] M. Usama, H. Anwar, M. M. Shahid, A. Anwar, S. Anwar, and H. Hlavacs, "Vehicle and license plate recognition with novel dataset for toll collection," *arXiv preprint arXiv:2202.05631*, 2022.
- [18] Y. Chang, H. Hsu, J. Lee, C. Chueh, C. Chang, and L. Chen, "License plate character recognition using block-binary-pixel-sum features," in *International Conference on Computer, Networks and Communication Engineering (ICCNCE 2013)*. Atlantis Press, 2013, pp. 111–113.
- [19] G. Lalitha and S. Gopinathan, "An enhanced algorithm for segmenting and slicing license plate characters via cc analysis," *Journal of Education: Rabindra Bharathi University*, 2021.
- [20] S. Ktata, T. Khadhraoui, F. Benzarti, and H. Amiri, "Tunisian license plate number recognition," *Procedia Computer Science*, vol. 73, pp. 312–319, 2015.
- [21] M. S. Al-Shemarry, Y. Li, and S. Abdulla, "An efficient texture descriptor for the detection of license plates from vehicle images in difficult conditions," *IEEE transactions on intelligent transportation systems*, vol. 21, no. 2, pp. 553–564, 2019.

Genetic Algorithms Applied to the Searching of the Optimal Path in Image-based Robotic Navigation Environments

Fernando Martínez Santa, Fredy H. Martínez Sarmiento, Holman Montiel Ariza
Universidad Distrital Francisco José de Caldas
Bogotá, Colombia

Abstract—This paper describes an optimal-path finding strategy based on Genetic Algorithms, applied to mobile robots in static navigation environments. This strategy starts from an image or plan of the environment and is supported by some different image processing algorithms, mainly the image skeletonization. Three different strategies were tested, changing the domain of the optimization target function for the Genetic Algorithm, the first domain was all the points of the environment image less the obstacles or walls, the second domain was similar but using an image with the obstacles dilated, and the final domain was only the points of the skeleton image. The last tested domain is from 99.4% to 99.6% smaller than the others, that implied reductions from 95% to 96% in the overall execution time of the strategy. Likewise, three skeletonization algorithms were tested in order to use the one with less execution time in this proposal. Finally, the proposed path planning strategy was tested on the same environment changing the initial and final points giving as result a valid and optimized path for the mobile robot in all the tested cases, and an overall average optimization time less than 2 minutes. This last, validates this proposal for robotic navigation applications with static obstacles.

Keywords—Genetic algorithms; optimal path; optimization; robotic environment; mobile robots, image skeletonization

I. INTRODUCTION

Although robotics has been one of the areas with the most ongoing research over the years, today, it continues to expand its fields of application due to the incorporation of automation in daily life. One example of this is mobile robots, which are now looking to help people in daily tasks; this situation implies that mobile robots must be capable of a path in different environmental conditions or grounds. Hence, navigation has become an important robotics process, allowing a mobile robot to know its location and plan and follow a path, preventing collision with obstacles.

According to recent studies [1], there are some approaches to the solution of the navigation problem in robotics based on deterministic and non-deterministic algorithms, which focus on optimal path planning [2] and collision prevention [3]. When reviewing from the path planning point of view, where a route is planned in a given environment reducing the total cost associated with the trajectory, it is necessary to consider whether local path planning [4] or global path planning [5], [6] is being performed. Thus, different methods have been developed, classified into classical approaches and heuristic [1] or reactive approaches [7]. In recent years, heuristic or reactive approaches have been the most used due to their robustness

to handle the uncertainty present in the environment and real-time navigation problems. Some heuristic or reactive methods that have been used are Genetic Algorithm (GA) [8], [9], Ant Colony Optimization (ACO) [10], [11], Particle Swarm Optimization (PSO) [12], [13], [14], [15], Neural Networks (NN) [16], Fuzzy Logic (FL) [17], Dijkstra algorithm [18], A* algorithm [19], among others.

Genetic algorithms (GAs) are part of evolutionary computing, one of Artificial Intelligence techniques that exist today. GA is a meta-heuristic method for solving searching and optimization problems, where a new population is generated from the fitness value of the previous generation. It is based on the phenomenon of natural selection and genetic operations such as mutation and crossover [20]. Some studies have shown that GA is a robust search method that requires little information about the environment to achieve reducing path length and producing smoother path for robot navigation, with some limitations in convergence rate and time-consuming process [21], [22].

Therefore, the aim of this research is to propose a path planning strategy for mobile robots based on digital image processing and Genetic Algorithms. The main idea is exploring the GA as selection and optimization tool for searching a valid and optimized path for a mobile robot in its environment, starting from an image or plan of that environment.

The remainder of this paper is organized as follows. In Section II, a brief description of the robot environment is given, likewise the description of the data flow (pipeline) of the proposed strategy; the pre-processing image operations for the environment image are presented in Sections II-A, II-B and II-C; Section II-D gives the first process of the path planning, the computing of the skeleton of the image. Section II-F describes the application of a GA for finding an optimal path. Section III shows the experiments and results and the Section IV gives our final conclusions.

II. METHODOLOGY

The path planning proposal for mobile robots shown in this document, is supported by digital image processing [23], [24] and it is based on an initial image of the navigation environment, which can be a plan of the room or a photo taken from above. For the scope of this article the initial images are as the one shown in Fig. 1 which represents a plan of a building floor where the mobile robot has to navigate. In that

initial image the floor walls are shown in *black* and the free-navigable space are shown in *white*, likewise the p_s represents the initial or starting point of the robot and p_e the final or ending point.

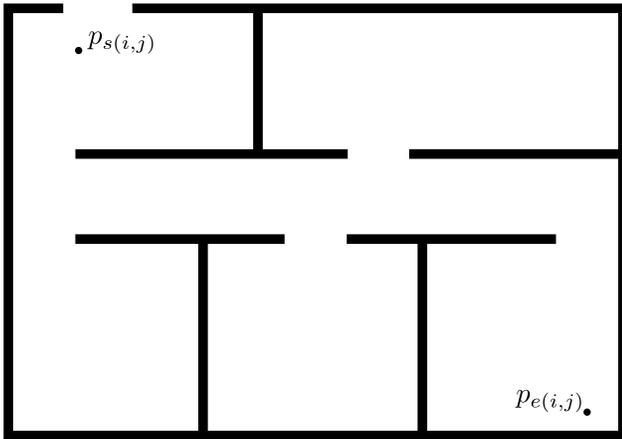


Fig. 1. Navigation Environment for the Mobile Robot.

The proposed path planning strategy is composed of the following steps: first, a resizing process is applied to the environment image in order to reduce and standardize the input. Second, the resized image is turned into binary to be compatible with the next steps of the process. After, two images are calculated, the first one is an image with the walls expanded (as wide as the radius of the robot), this image will be used to determine possible collisions in the next steps. The other image is the skeleton [25], [26] (medial axis) of the free-navigable space, this one is used for calculating the final path as far as possible to the obstacles (walls). Then, based on the skeleton of the environment image, the array of all the possible navigable points is stored. Finally, a short path is found using a Genetic Algorithm (GA) as optimizing and/or searching algorithm. The GA uses both the image skeleton and the walls dilated image to find the optimal path. The complete pipeline of proposed strategy is shown in the diagram of the Fig. 2. The overall path planning strategy was implemented by using *Python 3* programming language and mainly the *Scikit Image* module for the digital image processing operations. Next subsections show a detailed explanation of each of these steps.

A. Image scaling

In order to reduce the computing time of the overall path planning strategy, the images to work with have to be the small as possible, due to that, the resolution of the input image and therefore all the generated and used images is limited to 700k pixel, specifically images of 1000x700 pixels. This last is very important mainly for the GA, due to some image processing operations are part of the fitness function or target function, that implies it has to be iterated very much times until reach the convergence value. The exponential increasing of the execution time in image processing algorithms where the resolution of the input image increases, will imply a very high execution time for the overall strategy, for that reason is very important to limit the input image resolution. In the scaling or resizing operation that was applied to the input image, no anti-aliasing

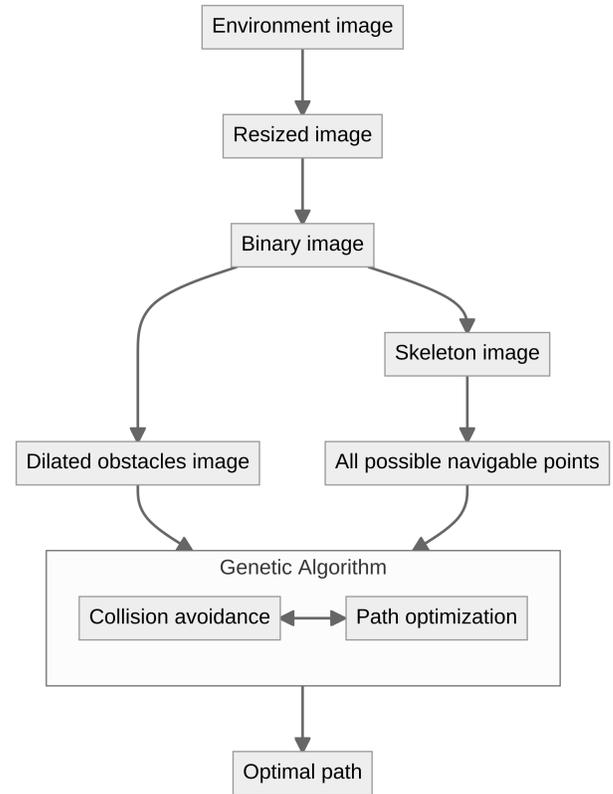


Fig. 2. Pipeline of the Proposed Path Planning Strategy.

method was used in order to not affect the original shape of the border objects in the environment.

B. Image Binarization

Once scaled, the image is normalized and binarized, this is achieved first by turning all the pixel values (bytes) of the image into normalized values from 0 to 1, by means of a simple product. After that, a threshold operation is performed following the eq. 1, where I and J are the maximum dimensions of the normalized image A and the output binary image B , likewise Th is the threshold defined as 0.5. At the end of this process, the image B has the obstacles in *black* (False) and the navigable space in *white* (True), as shown in Fig. 3.

$$\forall i \in I, j \in J : \begin{cases} A_{ij} > Th \rightarrow B_{ij} = True \\ A_{ij} \leq Th \rightarrow B_{ij} = False \end{cases} \quad (1)$$

C. Obstacle Dilation

In order to keep the robot to a safe distance from the obstacles or walls of the environment, that distance is calculated from to the maximum radius (measure from the center to the maximum distance to this one) of the robot following the eq. 2, where r_d is the dilation radius, r_m is the robot maximum radius and Δr is a radius tolerance defined as 10%. The resultant r_d is represented in pixel units and it is rounded to the floor.

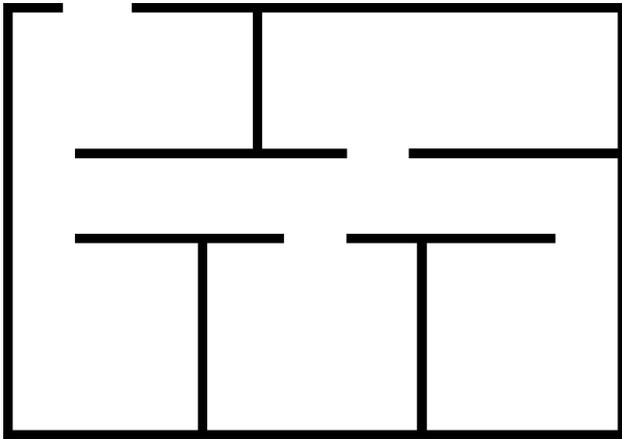


Fig. 3. Resized and Binarized Image of the Environment.

$$r_d = \lfloor r_m + \Delta r \rfloor \quad (2)$$

In a binary image, the dilation process is achieved over the *white* objects, so it is necessary to invert the image to obtain the desired *white* obstacles. At the end, the obtained image is inverted again to obtain an image with the dilated *black* obstacles. After inverting the obstacles binary image, a morphology dilation operation is performed by means of using a 2D convolution between that image and a disk shape a radius equivalent to r_d . After that, the image is inverted back, and the result is shown in Fig. 4, where the area of the obstacles or walls are expanded because of the dilation operation. All of this, pretends to avoid collisions between the robot with the walls due to the maximum radius of the robot r_m was taken into account.

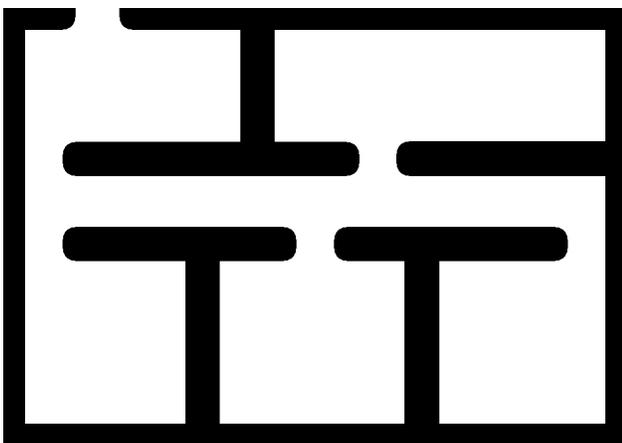


Fig. 4. Dilated Obstacles Image.

D. Skeletonization Algorithm

The image skeletonization algorithms pretend to find a medial axis of the shapes on a binary image applying an iterated and controlled image erosion operation until obtained a thin line [27]. This obtained line represents the medial axis of each object in the image. There are some different proposed

algorithms about such as the one proposed by Zhang et al. [28], the Lee's proposal (et al.) [29], and other medial axis operation. For the scope of this paper, three different skeletonization algorithms were tested: the standard skeletonization (Zhang), the Lee's skeletonization and a standard medial axis obtaining algorithm. The main aim is to recognize which of them is the fastest in order to be applied in this proposal. The Fig. 5 shows the resultant execution times of the tested skeletonization algorithms. The tests was applied over an 1000×700 image that contains a possible navigation environment for a mobile robot. A total of 5 tests were done and the average execution time of each algorithm is the shown in Fig. 5. Then, according to those results, for this proposal only the Zhang skeletonization algorithm is applied for obtaining the medial axis points of the environment image, as reference for the navigation.

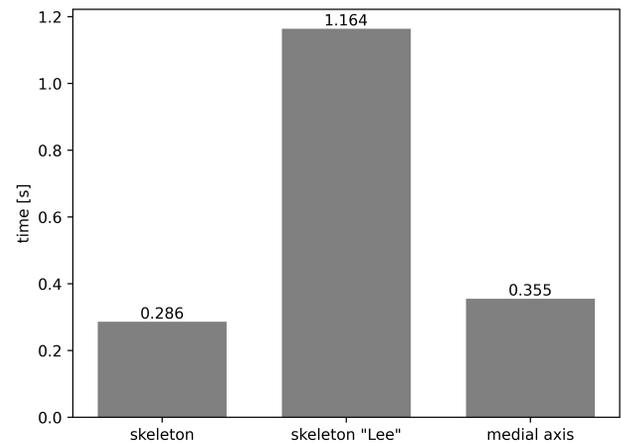


Fig. 5. Skeletonization Algorithms Execution Times.

The Zhang's skeletonization operation is applied to the resized environment image in order to obtain a thin line in the middle distance between obstacle or walls [30]. The Fig. 6 shows the resultant skeleton (in white) of the environment image, likewise the Fig. 7 shows original environment image plus the inverted skeleton image. This last image was obtained by means of applying logical *not* and logical *and* operations. As shown in the Fig. 7, the skeleton corresponds to the middle distance (medial axis) between halls walls and rooms walls, so it is a good reference for a free-collision navigation of the mobile robot.

E. Navigable Points

Once the environment image skeleton is obtained, all the points of the medial axis E_{ij} are stored in a bi-dimensional array P where each point $P_{k(x,y)}$ is a possible point of the final path. The storage operation of the skeleton set E starts from the skeleton image (see Fig. 6) and looks for the pixels in *white* (True), as the eq. 3 summarizes.

$$\forall k, i \in I, j \in J : E_{ij} = True \rightarrow P_k = (i, j) \quad (3)$$

F. Genetic Algorithm

For calculating and optimizing the path from the navigable points array, Genetic Algorithms (GAs) are used as

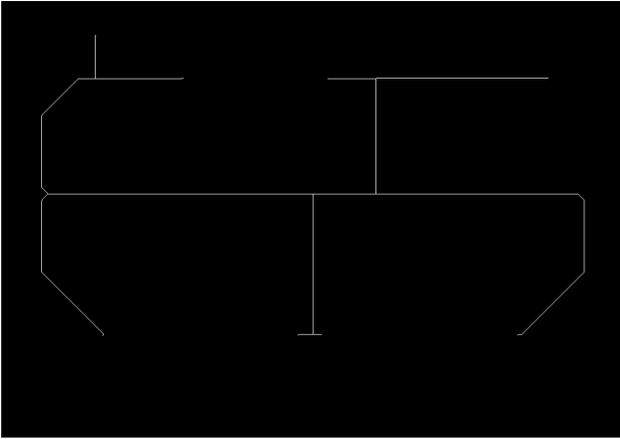


Fig. 6. Skeleton of the Environment Image.

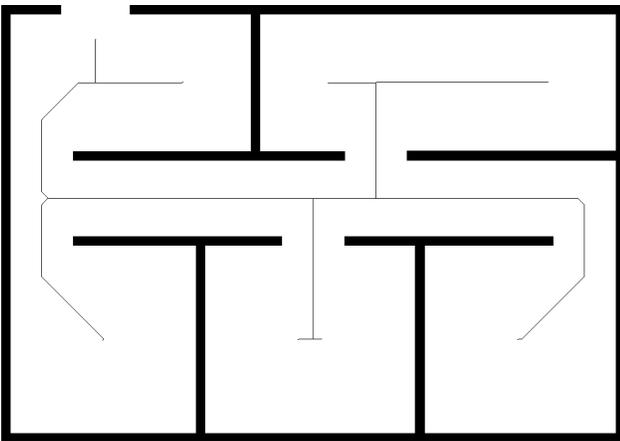


Fig. 7. Image of the Navigation Environment Plus its Skeleton.

optimization/searching tool [20] in order to find the shortest path using as reference the navigable points array obtained from the environment skeleton image, and at the same time avoiding collisions with obstacles and walls using the dilated obstacles image previously mentioned.

1) *Collision Avoidance*: The avoidance of possible collisions between each segment of the final path and the obstacles or walls is achieved applying some binary image operations, specifically a binary exclusive disjunction *XOR* between the dilated walls image and the same image with the specific segment drawn in *white* as shown in Fig. 8, where p_1 and p_2 are the points of the path segment. When there is a collision, a white segment line will be down over an obstacle or wall, producing that the two images are different. The eq. 4 has to be accomplished by all the pixels of both images for validating the non-collision condition. Where again, I and J are the valid set of indices in the dilated obstacles image D and the copy of the same image plus the drawn line L .

$$\forall i \in I, j \in J : D_{ij} \leftrightarrow L_{ij} = False \quad (4)$$

2) *Fitness Function*: For starting the searching and optimization process, the starting point p_s and the ending point p_e

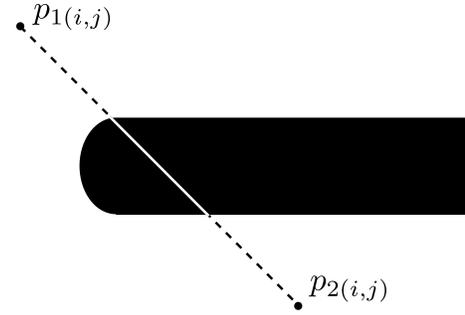


Fig. 8. Drawn Line for Testing the Collision with the Obstacles (walls).

are appended to the navigable points array P . Once completed the navigable points array, it is necessary to define the *fitness function* or *target function* to optimize $f(X)$, which depends on the solution set X shown in eq. 5.

$$X = \{x_0, x_1, x_2 \dots x_n\} \quad (5)$$

where each x represents a reading index of the navigable points set P which is based on the skeleton of the environment image. Then, the *fitness function* depends on the summation of the distances of each segment $\overline{p_i p_{(i+1)}}$ from p_s to p_e , that also accomplishes the eq. 4, the complete definition is shown in eq. 6.

$$f(X) = \sum_{i=0}^n |\overrightarrow{p_i p_{(i+1)}}| + d_c \quad (6)$$

Listing 1: *fitness function* Implementation on Python 3.

```
def obj_function(X):
    d = 0

    route_x ,
    route_y = build_route( X,
                           start_p ,
                           end_p )

    for i in range(len(route_x)-1):

        dx = route_x[i] - route_x[i+1]
        dy = route_y[i] - route_y[i+1]
        d += np.sqrt( dx**2 + dy**2 )

        imt = (imd == True)
        rr , cc = line( route_y[i],
                       route_x[i],
                       route_y[i+1],
                       route_x[i+1] )
        imt[rr , cc] = True

    imc = imt == imd
    if sum(sum(imc)) != imd.size:
        d += (max_i + max_j)
```

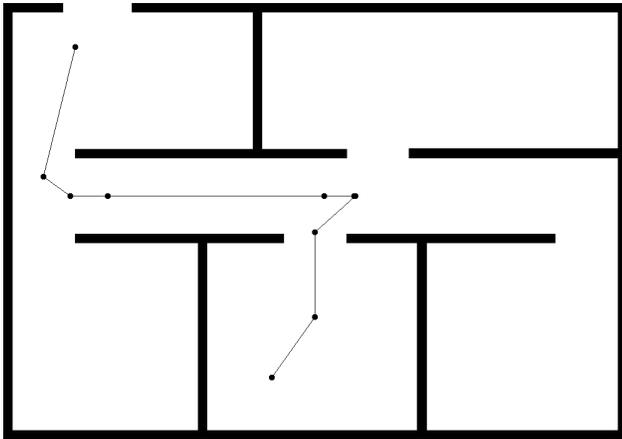



Fig. 10. Resultant Optimized Image, Case 2.

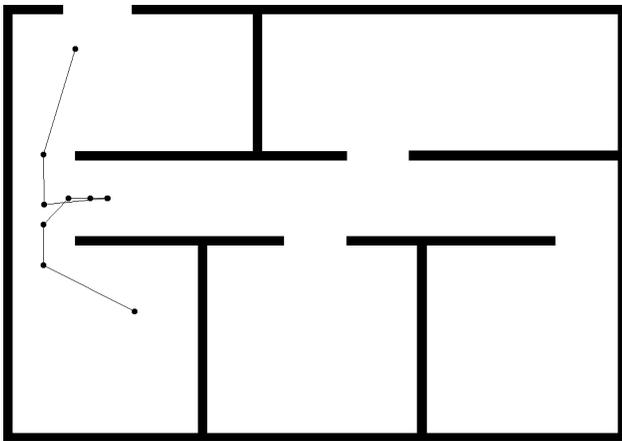


Fig. 11. Resultant Optimized Image, Case 3.

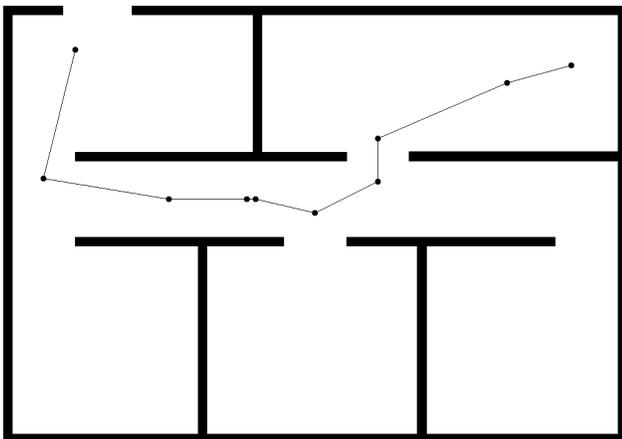


Fig. 12. Resultant Optimized Image, Case 4.

As future work, an adaptable and automatic change of the total number of points of the path is proposed in order improve the convergence times and the optimal result when the starting and ending points are close as the example shown in Fig. 11.

ACKNOWLEDGMENT

This work was supported by Universidad Distrital Francisco José de Caldas, specifically by the Technology Faculty. The views expressed in this article are not necessarily endorsed by Universidad Distrital. The authors thank the ARMOS research group for the given technical support and resources.

REFERENCES

- [1] R. Manzoor and N. Kumar, "Mobile robot path planning approaches: Recent developments," in *Innovations in Information and Communication Technologies (IICT-2020)*, P. K. Singh, Z. Polkowski, S. Tanwar, S. K. Pandey, G. Matei, and D. Pirvu, Eds. Cham: Springer International Publishing, 2021, pp. 301–308.
- [2] M. N. Zafar and J. Mohanta, "Methodology for path planning and optimization of mobile robots: A review," *Procedia computer science*, vol. 133, pp. 141–152, 2018.
- [3] L. Blasi, E. D'Amato, M. Mattei, and I. Notaro, "Path planning and real-time collision avoidance based on the essential visibility graph," *Applied Sciences*, vol. 10, no. 16, p. 5613, 2020.
- [4] J. Krejsa and S. Vechet, "Determination of optimal local path for mobile robot," in *Mechatronics 2017*, T. Březina and R. Jabłoński, Eds. Cham: Springer International Publishing, 2018, pp. 637–643.
- [5] T. T. Mac, C. Copot, D. T. Tran, and R. De Keyser, "A hierarchical global path planning approach for mobile robots based on multi-objective particle swarm optimization," *Applied Soft Computing*, vol. 59, pp. 68–76, 2017.
- [6] F. M. Santa, S. O. Rivera, and M. A. Saavedra, "Enfoque de navegación global para un robot asistente," *Tecnura*, vol. 21, no. 51, p. 105, 2017.
- [7] B. Patle, A. Pandey, D. Parhi, A. Jagadeesh *et al.*, "A review: On path planning strategies for navigation of mobile robot," *Defence Technology*, vol. 15, no. 4, pp. 582–606, 2019.
- [8] R. M. C. Santiago, A. L. De Ocampo, A. T. Ubando, A. A. Bandala, and E. P. Dadios, "Path planning for mobile robots using genetic algorithm and probabilistic roadmap," in *2017IEEE 9th international conference on humanoid, nanotechnology, information technology, communication and control, environment and management (HNICEM)*. IEEE, 2017, pp. 1–5.
- [9] X. Liang, P. Jiang, and H. Zhu, "Path planning for unmanned surface vehicle with dubins curve based on ga," in *2020 Chinese Automation Congress (CAC)*. IEEE, 2020, pp. 5149–5154.
- [10] M. Brand, M. Masuda, N. Wehner, and X.-H. Yu, "Ant colony optimization algorithm for robot path planning," in *2010 international conference on computer design and applications*, vol. 3. IEEE, 2010, pp. V3–436.
- [11] K. Akka and F. Khaber, "Mobile robot path planning using an improved ant colony optimization," *International Journal of Advanced Robotic Systems*, vol. 15, no. 3, p. 1729881418774673, 2018.
- [12] K. Su, Y. Wang, and X. Hu, "Robot path planning based on random coding particle swarm optimization," *International journal of advanced computer science and applications*, vol. 6, no. 4, pp. 58–64, 2015.
- [13] H. Mo and L. Xu, "Research of biogeography particle swarm optimization for robot path planning," *Neurocomputing*, vol. 148, pp. 91–99, 2015.
- [14] X. Li, D. Wu, J. He, M. Bashir, and M. Liping, "An improved method of particle swarm optimization for path planning of mobile robot," *Journal of Control Science and Engineering*, vol. 2020, 2020.
- [15] A. Tharwat, M. Elhoseny, A. E. Hassanien, T. Gabel, and A. Kumar, "Intelligent bézier curve-based path planning model using chaotic particle swarm optimization algorithm," *Cluster Computing*, vol. 22, no. 2, pp. 4745–4766, 2019.
- [16] H.-y. Zhang, W.-m. Lin, and A.-x. Chen, "Path planning for the mobile robot: A review," *Symmetry*, vol. 10, no. 10, p. 450, 2018.
- [17] A. Pandey and D. R. Parhi, "Optimum path planning of mobile robot in unknown static and dynamic environments using fuzzy-wind driven optimization algorithm," *Defence Technology*, vol. 13, no. 1, pp. 47–58, 2017.

- [18] S. A. Zanlongo, L. Bobadilla, and Y. T. Tan, "Path-planning of miniature rovers for inspection of the hanford high-level waste double shell tanks," in *Florida Conference on Recent Advances in Robotics (FCRAR)*, 2017.
- [19] J. D. Contreras, F. Martínez *et al.*, "Path planning for mobile robots based on visibility graphs and a* algorithm," in *Seventh International Conference on Digital Image Processing (ICDIP 2015)*, vol. 9631. SPIE, 2015, pp. 345–350.
- [20] S. B. Mane and S. Vhanale, "Genetic algorithm approach for obstacle avoidance and path optimization of mobile robot," in *Computing, Communication and Signal Processing*, B. Iyer, S. Nalbalwar, and N. P. Pathak, Eds. Singapore: Springer Singapore, 2019, pp. 649–659.
- [21] N. Adzhar, Y. Yusof, and M. A. Ahmad, "A Review on Autonomous Mobile Robot Path Planning Algorithms," *Advances in Science, Technology and Engineering Systems Journal*, vol. 5, no. 3, pp. 236–240, 2020.
- [22] C. Lamini, S. Benhlima, and A. Elbekri, "Genetic algorithm based approach for autonomous mobile robot path planning," *Procedia Computer Science*, vol. 127, pp. 180–189, 2018.
- [23] N. Roy, R. Chattopadhyay, A. Mukherjee, and A. Bhuiya, "Implementation of image processing and reinforcement learning in path planning of mobile robots," *International Journal of Engineering Science*, vol. 15211, 2017.
- [24] S. Luo, Y. Singh, H. Yang, J. H. Bae, J. E. Dietz, X. Diao, and B.-C. Min, "Image processing and model-based spill coverage path planning for unmanned surface vehicles," in *OCEANS 2019 MTS/IEEE SEATTLE*. IEEE, 2019, pp. 1–9.
- [25] D. H. Ko, A. U. Hassan, S. Majeed, and J. Choi, "Skelgan: A font image skeletonization method," *Journal of Information Processing Systems*, vol. 17, no. 1, pp. 1–13, 2021.
- [26] X. Bai, L. Ye, J. Zhu, L. Zhu, and T. Komura, "Skeleton filter: A self-symmetric filter for skeletonization in noisy text images," *IEEE Transactions on Image Processing*, vol. 29, pp. 1815–1826, 2019.
- [27] M. Nazarkevych, S. Dmytruk, V. Hrytsyk, O. Vozna, A. Kuza, O. Shevchuk, Y. Voznyi, I. Maslanych, and V. Sheketa, "Evaluation of the effectiveness of different image skeletonization methods in biometric security systems," *International Journal of Sensors Wireless Communications and Control*, vol. 11, no. 5, pp. 542–552, 2021.
- [28] T. Y. Zhang and C. Y. Suen, "A fast parallel algorithm for thinning digital patterns," *Communications of the ACM*, vol. 27, no. 3, pp. 236–239, 1984.
- [29] T.-C. Lee, R. L. Kashyap, and C.-N. Chu, "Building skeleton models via 3-d medial surface axis thinning algorithms," *CVGIP: Graphical Models and Image Processing*, vol. 56, no. 6, pp. 462–478, 1994.
- [30] F. M. Santa, E. J. Gomez, and H. M. Ariza, "Global path planning for mobile robots using image skeletonization," *Indian J. Sci. Technol.*, vol. 10, p. 14, 2017.

Anomaly Detection using Network Metadata

Khaled Mutmbak¹, Sultan Alotaibi², Khalid Alharbi³, Umar Albalawi⁴, Osama Younes⁵

College of Computing and Information Technology, University of Tabuk, Tabuk, 71491, Saudi Arabia^{1,2,3,4,5}
Faculty of Computers and Information, Menoufia University, Menoufia 32951, Egypt⁵

Abstract—The proliferation of numerous network function today gave rise to the importance of network traffic classification against various cyber-attacks. Automatic training with a huge number of representative data necessitates the creation of a model for an efficient classifier. As a result, automatic categorization requires using training techniques capable of assigning classes to data objects based on the activities supplied to learn classes. Predefined classes allow for the detection of new items. However, the analysis and categorization of data activity in intrusion detection systems are vulnerable to a wide range of threats. Thus, New methods of analysis must be developed in order to establish an appropriate approach for monitoring circulating traffic in order to solve this problem. The major goal of this research is to develop and verify a heterogeneous traffic classifier that can classify the collected metadata of networks. In this study, a new model is proposed, which is based on machine learning technique, to increase the accuracy of prediction. Prior to the analysis stage, the gathered traffic is subjected to preprocessing. This paper aims to provide the mathematical validation of a novel machine learning classifier for heterogeneous traffic and anomaly detection.

Keywords—Anomaly detection; network metadata; packet analysis; intrusion detection system; machine learning; classification; heterogeneous traffic

I. INTRODUCTION

As part of network forensics, network traffic and event logs are commonly referred to as being sniffed, recorded, acquired, and analyzed to investigate a network security incident. It enables the investigator to study network traffic and records to identify and locate the assaulting system. Computers, smart-phones, tablets, and other network-connected devices continue to grow. As the frequency of assaults against networked systems increases, the criticality of network forensics grows. Most previous studies confront two fundamental issues in extracting external and internal data, making traffic flow prediction a difficult endeavour. Currently, available solutions do not completely use the fundamental properties of short-term nearby and long-term periodic temporal patterns in terms of their various roles. In terms of the extrinsic task, current work has primarily used hand-crafted fusion algorithms to incorporate external inputs, however, there are still challenges with generalization [1].

The examination of a traffic incident is divided into two stages: Appearance check is the initial stage in the process of determining the Bloom filter (period) including an excerpt. To find the flows that conveyed the excerpt, the second phase is termed "flow determination," and it involves combining the excerpt blocks with the flows found by the Bloom filter. It

was key difficulties handled by HBF [2], such as ensuring that blocks were aligned and that they were consecutively placed.

Cybercrime is a constant danger to computer networks. No security mechanism can guarantee complete safety. Even the most advanced network security measures are unable to identify and prevent all assaults, particularly those that are new and unknown. In certain circumstances, preventing cybercrime is impossible. Suppose that confidential information about a company is leaked over its network. How can security specialists track down cybercriminals? Let us consider the following scenario: an organization's internal network has been infected by a worm, and the organization's Intrusion Detection and Prevention System (IDS/IPS) was unable to identify and block the worm's dissemination. How can you track down the person who spreads the virus or the afflicted systems? As a result, in addition to preventative security systems, tools and methodologies for investigating cybercrime after it has occurred are required. This is the function of network forensics and the tools that it provides [2].

Recording and storing raw network traffic is the most basic method of network research. Traffic recording makes it feasible to examine any networking event that occurs. It is possible to scan through the recorded traffic for the leaked information or the worm's signature to determine where it originated and where it ended up. "Attribution" is the term for this operation. The most difficult challenge with this system is the exceedingly costly storing of large amounts of data [3]. In addition, the invasion of privacy is a concern with traffic recording. By monitoring network traffic, it is possible to gain access to the personal information of users. As a result of the increasing difficulty in providing both privacy and network forensics, new Internet designs and protocols have been proposed [4]. However, implementing such modifications would be prohibitively expensive, making them impractical in practice.

In the field of traffic categorization, three groups of methodologies exist port-based, payload-based, and machine learning-based methods [5]. The identification of network traffic based on port numbers is a straightforward process that depends on mapping programs to well-known port numbers. Regrettably, port-based categorization algorithms have grown erroneous as a result of the increased use of dynamic port numbers by numerous apps. Payload-based approaches necessitate the analysis of the payload of each packet. Privacy regulations and encryption, on the other hand, may prevent traffic payloads from being accessed. As a result of this, deep packet inspection (DPI) is expensive in terms of both computation and signature maintenance [6].

II. MOTIVATION

Machine learning-based solutions have the potential to overcome some of the restrictions associated with port- and payload-based systems. More precisely, machine learning approaches can classify Internet traffic based on application-neutral traffic data. When it comes to how long it takes to send and receive a particular message, there are several variables that may be taken into consideration. Furthermore, it has the potential to minimize computing costs while also making it easier to identify encrypted traffic.

There are two main applications for network forensics. The first, which focuses on network security, is keeping an eye out for unusual traffic patterns and spotting breaches. On a hacked system, an attacker may be able to delete all log files. Consequently, network-based evidence may be the sole evidence accessible for forensic investigation in this situation. Law enforcement can also take advantage of network forensics by interpreting human communication represented through e-mails or other forms of electronic correspondence and reassembling transmitted information, looking for keywords, and so on [7].

Today's world is evolving at a rapid pace, and the internet is critical for quicker communication between people or machines, faster transactions, and faster fulfillment of duties (tasks). However, the internet is also a major victim of cybercrime. Transactions over the internet are the main draw for attackers. To do this, we need a forensic technology known as "Network Metadata" to help us identify the perpetrators of cybercrime and their methods of attack. Network Metadata is a sub-field of digital forensics research that deals with computer networks. The collection of network traces from the victim system for examination is a common practice in network forensics, whether the crime has been discovered or after it has been committed. The evidence gathered can be used to bring the perpetrator to justice in a criminal court of law. While digital forensics involves the examination of static data, network forensics involves the examination of volatile and dynamic data [8].

III. RELATED WORK

The study [9] establishes a network intrusion criminal system based on the switching scheme (NIFSTC) that may detect criminality in networked situations and identify digital evidence automatically. The advantage of NIFSTC is that it does not require a standard forensic network to be built, hence it has superior detection performance in practice than traditional approaches. For the most modern network forensic methodologies, the KDD Cup Experiment Series 1999 dataset shows NIFSTC's highest true positive (TP) and lowest rate false positive (FP) .

The authors [10] introduced SPIE (Source Path Isolation Engine) in this regard, which calculates the first eight bytes of the payload and packet digests (i.e. hashes) from the header. A brief period of time is allowed for the digestion of these digests in a bloom filter. If a third-party device, such as an IDS or a firewall, identifies suspicious activity, SPIE can be used to track down the source of a packet.

In the research [11], the focus is on the security risks of the botnet through which DDoS attacks, worms and spam attacks

are implemented. For network security forensic investigation, the researchers recommended the design and implementation of a cloud-based security center. Also, cloud storage is used to store the acquired traffic data, which is then processed utilizing cloud computing.

A tool that explores the architecture of the network forensic is proposed in [12], which is called NetFo (Network Forensic) analysis tool. It captures packets using Winpcap technology and It can be used as a monitoring and management tool. NetFo can discover session information, keywords, bookmarks, hostnames, IP addresses, and other information.

As explained in [13], due to many requirements that were not addressed in this design space, developing a forensic network architecture is a complex task.

The authors [14] present a real-life case study in which they reconstruct a crime scene in relation to a victim's previous Facebook session using digital evidence collected and analysed via access to a desktop computer's RAM, with a focus on some distinct chains that could be used to reconstruct a previous Facebook session.

Huaxin et al. [15] developed a framework for extracting four types of characteristics from real-world Wi-Fi data, as well as supervised machine learning approaches for estimating user demographics. The study was based on Wi-Fi traffic information from 28,158 users during a five-month period. According to the testing results, the best accuracy in predicting gender and education level is 82% and 78%, respectively. Users' demographics may be predicted with a precision of 69% and 76% utilizing HTTPS traffic, even in encrypted transmission (i.e., across the internet). Being forensically prepared increases the degree of security in both cloud and on-premises computing. As a result, research in the fields of cloud and network security may also apply to IoT-centered forensics investigations. After all, traditional computer networks and Internet of Things (IoT) networks are also vulnerable to security flaws. Because IoT systems interact with the physical environment more frequently than traditional systems, they are susceptible to a greater number of physical and digital dangers. As a result, the work introduced in [16] was dedicated for securing the IoT domain.

The authors in [17] provided an overview of forensic advancements related to the IoT as well as the remaining hurdles. They focused on the taxonomy and criteria in the IoT Forensics. However, they did not discuss historical and current frameworks, standardization and certification difficulties in the IoT Forensics.

IV. THE PROPOSED MODEL

The ML models are becoming widespread in recent years because of mitigating a variety of complex relationships and acquiring the most favorable solutions by general evolution. The ML models have the ability to discover nonlinear relationships and complex functions among independent and dependent variables based on processing and classifying the data through training. An ML technique is comprised of algorithms with many models based on artificial intelligence. In this paper, five ML classifiers are used and compared in terms of the highest accuracy.

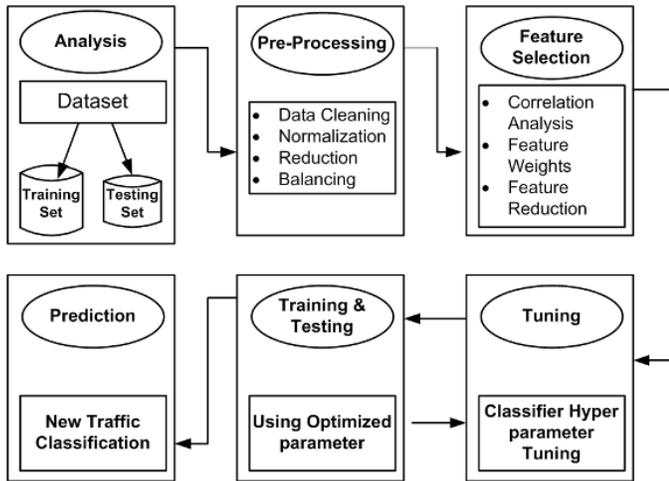


Fig. 1. The Proposed Model based on Machine Learning

Fig. 1 depicts the overall layout of the proposed framework based on machine learning approaches for network anomaly detection. It represents the phases that the model goes through and includes a large number of distinct processes. In the first phase, the dataset is analyzed and split into training and testing sets. For both training and testing, the attribute vectors are sliced in a 70:30 ratio. Next, in the pre-processing phase, the dataset is cleaned, features containing categorical data are normalized, and records including incorrect data are removed. Following, in the feature selection phase, the features are analyzed according to their weights and we choose most important features to define the attacks. Next, in the tuning phase, parameters of chosen classifier are tuned and optimized using a grid search. At the end, the optimized classifier is used for training and testing datasets, which are used for prediction of new traffic records.

A. Dataset

Data are the most valuable asset to develop an efficient intrusion detection system. CICIDS2017 [18] is the most recent intrusion detection evaluation dataset. It was created by the Canadian Institute for Cybersecurity at the University of New Brunswick. The CICIDS2017 dataset was constructed using the Network Traffic Flow analyzer. It was captured over a duration of 5 days over which 83 features and 15 classes were captured [19]. One of these classes represents the normal network traffic (defined as Benign) while the other 14 represent anomaly traffic (called Attacks). The names and numbers of these classes are shown in Table I. Compared to older and traditional datasets, such as KDD-99 [20], DARPA 98/99 [21] and ISCX2012 [22], CICIDS2017 dataset has the following advantages:

- Cover the current trends of attacks.
- Represent real-world data.
- Datasets are labelled.
- Attacks based on many protocols are included, such as HTTP, HTTPS, FTP, SSH, and email protocols.

For these reasons, the CICIDS2017 dataset is selected.

TABLE I. NUMBER OF STREAM RECORDS FOR ATTACK TYPES IN THE CICIDS2017 DATASET

Attack Type	Records
BENIGN	2359087
DoS Hulk	231072
PortScan	158930
DDoS	41835
DoS GoldenEye	10293
FTP-Patator	7938
SSH-Patator	5897
DoS slowloris	5796
DoS Slowhttptest	5499
Bot	1966
Web Attack – Brute Force	1507
Web Attack – XSS	652
Infiltration	36
Web Attack – Sql Injection	21
Heartbleed	11

B. Data Pre-Processing

As explained in the last section, the CICIDS2017 dataset contains 3119345 stream records and 83 features containing 15 class labels (one for normal traffic and 14 for attacks). To ensure that the dataset is ready to be trained, we need to clean and normalize it.

As most of the datasets, CICIDS2017 dataset contains some undesirable elements that must be removed. In CICIDS2017 dataset, because the network traffic was collected using the CICFlowMeter tool, some flag features have constant values (0 or 1), such as “Bwd URG Flags” and “Bwd P SH Flags”. These features were removed from the dataset because they have no impact on model results and to decrease the memory footprint of the dataset. Next step in the preprocessing phase is removing records that have missing class label, missing information, and invalid values such as “NaN” or “Inf”. After examining these records, 288602 records were removed.

If the dataset used for training of a classifier or detector suffers from high class imbalance problem, the classifier biases towards the majority class. As a result, the classifier shows lower accuracy with higher false alarm. Unfortunately, CICIDS2017 data set is prone to high class imbalance, as shown in Table 1. Therefore, to avoid this problem, the normal traffic records have been down sampled. In addition, to improve prevalence ratio and reducing class imbalance issue, few minority classes have been merged, such as Web Attacks. Therefore, the new dataset was partitioned into 70% for training (1571510 records) and 30% for testing (471453) sets.

C. Feature Selection

The goals of feature selection are to identify and remove unneeded, irrelevant and redundant features from the dataset. This help reduce the complexity of the predictive model without compromising its accuracy. Feature selection helps define most important features for detecting attacks on the dataset. First using correlation test, some features are removed from the dataset to reduce its size and enhance the performance.

Fig. 2 shows the correlation matrix, which shows the value of the correlation of variables and features with each other, which negatively or positively affects them. A correlation

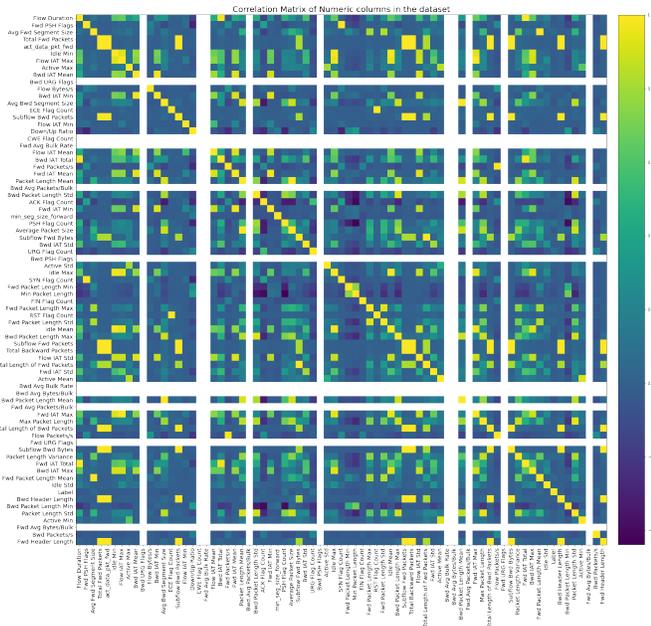


Fig. 2. Correlation Matrix

matrix is simply a table that displays the correlation coefficients for different variables. The matrix depicts the correlation between all the possible pairs of values in a table. It is a powerful tool to summarize a large dataset and identify and visualize patterns in the given data. Each cell in the table contains the correlation coefficient between the features that scales from 0 to 1. If the coefficient approaches 1, it means that it is more positive, meaning that both features have an impact on the prediction process, and whenever the value approaches 0, it means a negative correlation that does not benefit us in the process of prediction, and they have no effect.

By analyzing the correlation matrix, we found a strong correlation between the following features: (Fwd IAT Std, BwdIATMean), (Bwd Header Length, Fwd Header Length) and (Bwd Header Length, Subflow Fwd packets). Therefore, we delete the features that are not needed.

After removing correlated features, we still have large number of features. We need to use feature selection methods to determine the importance of a certain features in the detection of anomalous traffic. There are several feature selection methods in the literature, such as Fisher Score, T-Score, chi-squared tests, random forest, or regression. Using these five feature selection methods, each feature is given a weight of importance as to how useful they are. These weights of features are compared and sorted. Fig. 3 shows the most 10 important features that are used for training and testing in the proposed model.

V. RESULT AND DISCUSSION

In this section, the results will be presented and discussed based on the proposed machine learning techniques.

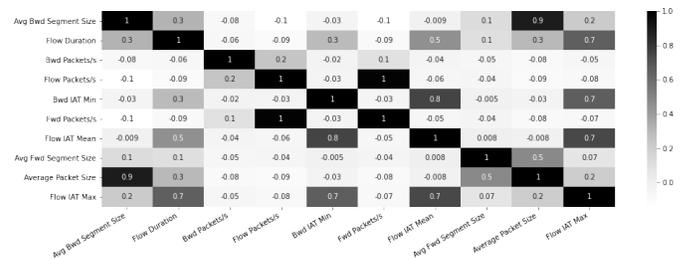


Fig. 3. Important Features

A. Data Classification Methods

In artificial intelligence, machine learning is regarded as a subfield. Automatic classification [[23], [24]] is one of interested subjects for machine learning. In order to handle classification difficulties, automatic learning employs a variety of methods that group homogenous classes of comparable data items together. In order to train the decision rule and develop a classifier, supervised learning is adopted. ML can be used to create a predictive model to detect unknown attacks in network traffic. However, one important problem in ML is to identify and select the most relevant feature characteristics, from which to build a specific model based on training data for a particular classification job [[24], [25],[26]].

Classification is a logical choice for doing predictions with discrete known outcomes when using a machine learning technique such as classification. Items are classified using a classification technique, which is a set of exact rules for categorizing objects based on the quantitative and qualitative factors that characterize the objects. There are a variety of goals for which data categorization is performed, the most prevalent of which is to assist with data security challenges, particularly in anomaly detection [[27], [28], [29]].

In this work, we adopted using five classifiers to categorize the network, which are: the Random Forest, Logistic Regression, Decision Tree Algorithm, SVM, and the k-nearest neighbors. The findings were then compared using performance metrics and classification reports. Through the optimization of the classifier, training and testing process are repeated, where the behavior of the classifier is changed until the intended behavior is accomplished.

B. Performance Measures

To evaluate the performance of the suggested classification methods for anomaly detection, we adopted the following measures: accuracy, recall, precision, and F1 Score. The confusion matrix is utilized to separate the prescient execution of the classification in the test data.

Fig. 4 shows a template for a binary confusion matrix that uses the four kinds of results: (true positives (TP), false negatives (FN), false positives (FP), and true negatives (TN)) along with the positive and negative classifications. The following measurement metrics are used to measure the performance of a dataset:

- 1) Accuracy calculates predicted observation ratios for

		Predicted condition	
		Positive (PP)	Negative (PN)
Total population = P + N			
Actual condition	Positive (P)	True positive (TP)	False negative (FN)
	Negative (N)	False positive (FP)	True negative (TN)

Fig. 4. Confusion Matrix

TABLE II. RESULT OF COMPARE BETWEEN CLASSIFIERS

Algorithm	Accuracy	Recall	Precision	F1 Score
K-Neighbors	95.84	97.51	95.84	97.84
Logistic regression	96.51	97.95	98.45	98.20
SVM	93.69	96.43	96.84	96.63
Decision Tree	97.11	98.31	98.69	98.50
Random Forest	98.63	98.82	99.80	99.31

total observations:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

- 2) Recall is the ability of the proposed model to detect the attacks. Recall can be calculated from the number of detected attacks rather than the number of actual attacks.

$$Recall = \frac{TP}{TP + FN} \quad (2)$$

- 3) Precision is the ratio of predicted positive to total positive observed predictions.

$$Precision = \frac{TP}{TP + FP} \quad (3)$$

- 4) F1 Score is the average of recall and precision values.

$$F1Score = 2 * \frac{Precision * Recall}{Precision + Recall} \quad (4)$$

C. Experimental Results

Table II and Fig. 5 show the results of applying five different machine learning techniques for classifying different types of attacks. Fig. 5 shows the confusion matrix for all algorithms. Table II shows the accuracy, precision, recall, F1 Score for each algorithm.

From the Table II, we can notice that the best algorithms are Random Forest and Decision Tree. This is because they have a high accuracy and precision rates. The worst algorithm is K-Neighbors because it had lower accuracy and precision rates.

VI. CONCLUSION AND FUTURE WORK

An intrusion detection system is an important protection tool for detecting complex network attacks. In this work we have developed a new model for network intrusion (anomaly) detection based on machine learning algorithms. The proposed

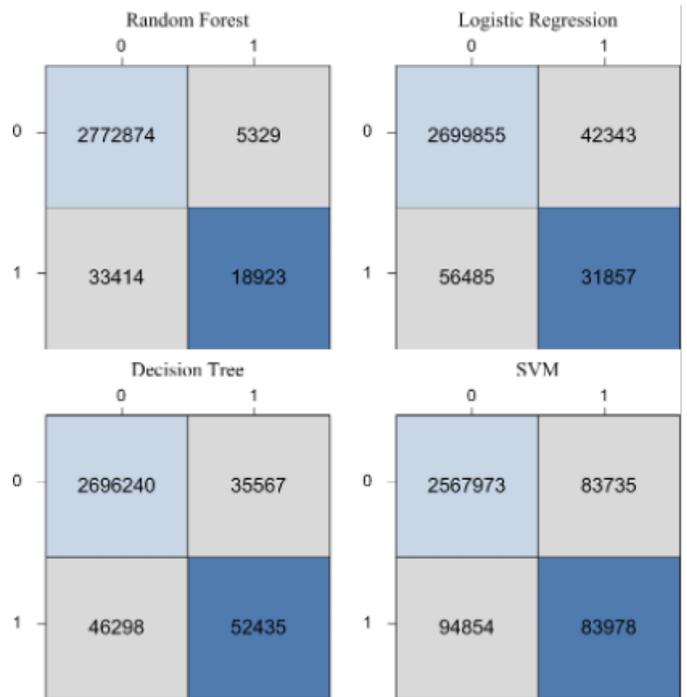


Fig. 5. Confusion Matrix for different Classification Methods

model consists of six phases: dataset analysis, pre-processing, feature selection, parameter tuning, training and testing. Using the proposed model, five machine learning algorithms have been investigated for classification of network anomaly detection, which are: K-neighbors, logistic regression, SVM, decision tree and random forest. The performances of these ML algorithms have been observed on the basis of their accuracy, recall, precision and F1 score. The dataset CICIDS-2017 has been used for training and testing, which consists of seven different types of attacks. According to results, compared to other ML algorithms, the performance of the random forest algorithm is better. This is because it has achieved the highest accuracy and precision rates for classification of anomaly detection, which are 98.63% and 99.80, respectively. Compared to related work, the performance of the proposed model is better. This is because of: (1) The dataset was carefully cleaned by removing noise and outlier data and solving imbalance issues. (2) The proposed feature selection technique removed correlated and irrelevant features from the dataset. (3) Parameters of chosen classifier are tuned and optimized using grid search. As a future work, we will investigate other machine learning and deep learning algorithms for network anomaly detection.

ACKNOWLEDGMENT

The authors would like to thank The University of Tabuk for providing research support and facilities.

REFERENCES

[1] S. Fang, X. Pan, S. Xiang, and C. Pan, "Meta-msnet: Meta-learning based multi-source data fusion for traffic flow prediction," *IEEE Signal Processing Letters*, vol. 28, pp. 6–10, 2020.

- [2] S. M. Hosseini, A. H. Jahangir, and M. Kazemi, "Digesting network traffic for forensic investigation using digital signal processing techniques," *IEEE Transactions on Information Forensics and Security*, vol. 14, no. 12, pp. 3312–3321, 2019.
- [3] D. Quick and K.-K. R. Choo, "Impacts of increasing volume of digital forensic data: A survey and future research challenges," *Digital Investigation*, vol. 11, no. 4, pp. 273–294, 2014.
- [4] M. Afanasyev, T. Kohno, J. Ma, N. Murphy, S. Savage, A. C. Snoeren, and G. M. Voelker, "Privacy-preserving network forensics," *Communications of the ACM*, vol. 54, no. 5, pp. 78–87, 2011.
- [5] T. T. Nguyen and G. Armitage, "A survey of techniques for internet traffic classification using machine learning," *IEEE communications surveys & tutorials*, vol. 10, no. 4, pp. 56–76, 2008.
- [6] W. Li and A. W. Moore, "A machine learning approach for efficient traffic classification," in *2007 15th International symposium on modeling, analysis, and simulation of computer and telecommunication systems*. IEEE, 2007, pp. 310–317.
- [7] R. Hunt and S. Zeadally, "Network forensics: an analysis of techniques, tools, and trends," *Computer*, vol. 45, no. 12, pp. 36–43, 2012.
- [8] G. Shrivastava, "Approaches of network forensic model for investigation," *International Journal of Forensic Engineering*, vol. 3, no. 3, pp. 195–215, 2017.
- [9] Z. Tian, W. Jiang, and Y. Li, "A transductive scheme based inference techniques for network forensic analysis," *China Communications*, vol. 12, no. 2, pp. 167–176, 2015.
- [10] A. C. Snoeren, C. Partridge, L. A. Sanchez, C. E. Jones, F. Tchakountio, S. T. Kent, and W. T. Strayer, "Hash-based ip traceback," *ACM SIGCOMM Computer Communication Review*, vol. 31, no. 4, pp. 3–14, 2001.
- [11] Z. Chen, F. Han, J. Cao, X. Jiang, and S. Chen, "Cloud computing-based forensic analysis for collaborative network security management system," *Tsinghua science and technology*, vol. 18, no. 1, pp. 40–50, 2013.
- [12] M. H. Mate and S. R. Kapse, "Network forensic tool-concept and architecture," in *2015 Fifth International Conference on Communication Systems and Network Technologies*. IEEE, 2015, pp. 711–713.
- [13] G. Shrivastava, "Network forensics: Methodical literature review," in *2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom)*. IEEE, 2016, pp. 2203–2208.
- [14] H.-C. Chu, D.-J. Deng, and J. H. Park, "Live data mining concerning social networking forensics based on a facebook session through aggregation of social data," *IEEE Journal on Selected Areas in Communications*, vol. 29, no. 7, pp. 1368–1376, 2011.
- [15] H. Li, H. Zhu, and D. Ma, "Demographic information inference through meta-data analysis of wi-fi traffic," *IEEE Transactions on Mobile Computing*, vol. 17, no. 5, pp. 1033–1047, 2017.
- [16] M. Stoyanova, Y. Nikoloudakis, S. Panagiotakis, E. Pallis, and E. K. Markakis, "A survey on the internet of things (iot) forensics: challenges, approaches, and open issues," *IEEE Communications Surveys & Tutorials*, vol. 22, no. 2, pp. 1191–1221, 2020.
- [17] I. Yaqoob, I. A. T. Hashem, A. Ahmed, S. A. Kazmi, and C. S. Hong, "Internet of things forensics: Recent advances, taxonomy, requirements, and open challenges," *Future Generation Computer Systems*, vol. 92, pp. 265–275, 2019.
- [18] I. Sharafaldin, A. H. Lashkari, and A. A. Ghorbani, "Toward generating a new intrusion detection dataset and intrusion traffic characterization." *ICISSp*, vol. 1, pp. 108–116, 2018.
- [19] World BankThe World Bank, "Intrusion detection evaluation dataset (cicids2017)," 2018, accessed April 2022, <http://www.unb.ca/cic/datasets/ids-2017.html>.
- [20] D. Dua and C. Graff, "UCI machine learning repository," 2017. [Online]. Available: <http://archive.ics.uci.edu/ml>
- [21] J. W. Haines, R. P. Lippmann, D. J. Fried, M. Zissman, and E. Tran, "1999 darpa intrusion detection evaluation: Design and procedures," MASSACHUSETTS INST OF TECH LEXINGTON LINCOLN LAB, Tech. Rep., 2001.
- [22] A. Shiravi, H. Shiravi, M. Tavallaee, and A. A. Ghorbani, "Toward developing a systematic approach to generate benchmark datasets for intrusion detection," *computers & security*, vol. 31, no. 3, pp. 357–374, 2012.
- [23] L. Igual and S. Seguí, "Introduction to data science," in *Introduction to Data Science*. Springer, 2017, pp. 1–4.
- [24] S. B. Kotsiantis, I. D. Zaharakis, and P. E. Pintelas, "Machine learning: a review of classification and combining techniques," *Artificial Intelligence Review*, vol. 26, no. 3, pp. 159–190, 2006.
- [25] V. Vapnik, "An overview of statistical learning theory," *IEEE Transactions on Neural Networks*, vol. 10, no. 5, 1999.
- [26] M. Rocha, P. Cortez, and J. Neves, "Evolution of neural networks for classification and regression," *Neurocomputing*, vol. 70, no. 16-18, pp. 2809–2816, 2007.
- [27] S. Hao, J. Long, and Y. Yang, "BI-ids: Detecting web attacks using bi-lstm model based on deep learning," in *International Conference on Security and Privacy in New Computing Environments*. Springer, 2019, pp. 551–563.
- [28] A. Verma, A. Singh *et al.*, "An approach to detect packets using packet sniffing," *International Journal of Computer Science and Engineering Survey*, vol. 4, no. 3, p. 21, 2013.
- [29] M. Idhammad, K. Afdel, and M. Belouch, "Detection system of http ddos attacks in a cloud environment based on information theoretic entropy and random forest," *Security and Communication Networks*, vol. 2018, 2018.

Correcting Arabic Soft Spelling Mistakes using BiLSTM-based Machine Learning

Gheith Abandah, Ashraf Suyyagh, Mohammed Z. Khedher
Department of Computer
Engineering
University of Jordan
Amman, Jordan 11942

Abstract—Soft spelling mistakes are a class of mistakes that is widespread among native Arabic speakers and foreign learners alike. Some of these mistakes are typographical in nature. They occur due to orthographic variations of some Arabic letters and the complex rules that dictate their correct usage. Many people forgo these rules, and given the identical phonetic sounds, they often confuse such letters. In this paper, we investigate how to use machine learning to correct such mistakes given that there are no sufficient datasets to train the correction models. Soft errors detection and correction is an active field in Arabic natural language processing. We generate training datasets using proposed transformed input approach and stochastic error injection approach. These approaches are applied to two acclaimed datasets that represent Classical Arabic and Modern Standard Arabic. We treat the problem as character-level, one-to-one sequence transcription problem. This one-to-one transcription of mistakes that include omissions and deletions is possible with adopted simple transformations. This approach permits using bidirectional long short-term memory (BiLSTM) models that are more effective to train compared to other alternatives such as encoder-decoder models. Based on investigating multiple alternatives, we recommend a configuration that has two BiLSTM layers, and is trained using the stochastic error injection approach with error injection rate of 40%. The best model corrects 96.4% of the injected errors and achieves a low character error rate of 1.28% on a real test set of soft spelling mistakes

Keywords—Arabic text; natural language processing; spelling mistakes; recurrent neural networks; bidirectional long short-term memory

I. INTRODUCTION

Arabic is one of the world's five major languages with over 290 million native speakers and a total of 422 million world speakers [1], [2]. Arabic is the fourth most common language on the Internet [3], [4] and the fastest growing language online [3]. A recent British Council report [5] ranks Arabic as the fourth needed language for the future based on economic and market factors, diplomatic and security priorities, mobility, tourism, and public interest.

The Arabic language has 28 letters which we show alongside their Unicode encoding in Table I. Of the 44 letters and diacritics listed, eight are letter variants: 0621–0626, 0629, and 0649.¹

Western linguists distinguish between two forms of standard Arabic: Classical Arabic (CA) and Modern Standard

Arabic (MSA). CA is the language of the Quran, ancient religious and liturgical texts, and old Arabic literature. MSA is the modern form that is syntactically, morphologically, and phonologically based on CA. MSA is the primary form of Arabic language used in education, business, media, news, and drafting laws and regulations. The contrast between MSA and CA “is mostly reflected in topic, vocabulary, and style rather than grammatical structure” [7].

Arabic is a diglossic language and Arabic linguists and speakers refer to both CA and MSA as *al-fusha* الفصحى to differentiate between the standard forms and the colloquial variants spoken throughout the Arab world. These dialects are not standardized and vary significantly socially and geographically. Furthermore, they are widely popular in use on social and messaging applications as well on the Internet [8].

The U.S. state department categorizes Arabic among the exceptionally difficult languages to learn for native English speakers [9]. Consequently, it is vital to develop and modernize automated tools that aid native speakers and learners alike in communicating in Arabic using correct grammar and spelling, especially in online communication. To this end, our previous work [10], [11], [12] investigated machine learning and hybrid approaches for Arabic text diacritization with recurrent neural networks (RNN). This novel work investigates Arabic text correction using RNN in contrast to traditional techniques.

Research efforts have normally focused on a subset of the typical spelling and writing mistakes encountered in Arabic texts; usually by addressing one or two at a time. We target in this work the most common type of spelling mistakes, which is the soft spelling mistakes.

A. Soft Spelling Mistakes

Soft spelling mistakes are a special type of lexical and semantic errors that are due to the orthographic variations of some Arabic letters. For example, at the beginning of a word, the Arabic letter *alef* - comes in bare *alef* and *alef with hamza* forms: *alef with madda above*, *alef with hamza above* and *alef with hamza below*. The former bare *alef* shape is called همزة وصل (*hamzaï wasl*) while the *alef* with *hamza* above or below is called همزة قطع (*hamzaï qat*). Replacing one form of *alef* with another is a major case of soft-spelling mistakes [13].

Despite having standard rules for the shape and placement of the *hamza* inside the word, or as known in Arabic الهمزة المتوسطة (*al-hamzaï al-mutawsiṭaï*); these rules are quite

¹The transliteration used in this paper is Romanization using Glosbe transliteration tools [6].

TABLE II. THE FREQUENCY OF SOFT SPELLING ERRORS ANALYZED IN AL-AMERI STUDY [14]

No.	Spelling Error Type	%	Wrong Spelling	Correct Spelling
1	Writing <i>middle hamza</i> on an <i>alef</i>	73%	قراءة (qirāʾ)	قراءة (qiraʾī)
2	Writing <i>alef maksura</i> instead of <i>alef</i>	71%	عفا (ʿafy)	عفا (ʿafa)
3	Writing <i>alef</i> instead of <i>alef maksura</i>	70%	بكا (baka)	بكي (baky)
4	Omitting <i>alef</i> following a <i>waw</i> at the verb end	67%	لا تنظرو (la tanzurw)	لا تنظروا (la tanzurwā)
5	Writing <i>teh</i> instead of <i>teh marbuta</i>	67%	كثرت (kathrat)	كثرة (katharī)
7	Writing <i>middle hamza</i> on <i>waw</i>	64%	المروؤة (al-murwʾat)	المروءة (al-murwʾai)
8	Writing <i>hamza</i> at the end of the word on <i>alef</i>	51%	أشياء (ʾashyāʾ)	أشياء (ʾashyaʾ)
9	Writing a standalone <i>hamza</i> at the word end	47%	خطء (khaʿʾ)	خطأ (khaʿʾā)
10	Writing <i>hamza</i> at the end of the word on <i>yeh</i>	47%	شيء (shayʾ)	شيء (shayʾ)
11	Dropping <i>lam</i> before a “solar letter”	38%	اسماء (asmaʾ)	السماء (al-sāmaʾ)
12	Writing <i>teh marbuta</i> instead of <i>teh</i>	37%	ساعة (saʿaʿ)	ساعات (saʿat)
13	Writing <i>middle hamza</i> on <i>yeh</i>	30%	يتفائل (yatafaʿal)	يتفائل (yatafaʿāʾl)
14	Writing <i>hamzaʾi qaṭʿ</i> instead of <i>hamzaʾi waṣl</i>	28%	ابن (abn)	ابن (abn)
15	Inserting <i>alef</i> after <i>waw</i> at the end of a word	25%	يبحثوا (yajithwā)	يبحثو (yajithw)
16	Confusing <i>teh marbuta</i> and <i>heh</i> at the word end	*	مكتبه (makatabh)	مكتبة (maktabaʾ)

* Common mistake yet not reported in this study

B. Approach and Contribution

In this paper, we propose using a tuned bidirectional long short-term memory (BiLSTM) recurrent neural network to detect and correct spelling mistakes written in either classical or modern standard Arabic. We target a subset of the most commonly encountered spelling mistakes in the Arabic language [14]. Mainly, errors in the soft misspelling category pertaining to *al-hamzat* (الهمزات), the different shapes of *alef*, and the common errors in shaping *teh* at the end of the word. We propose tackling the problem at the character-level and we propose letter conversion scheme that allows one-to-one training of the input sequences against the target sequences.

We propose and evaluate two approaches to train models to correct these mistakes. In the *transformed input* approach, the network is trained to predict correct spelling from transformed unified input. Whereas the *stochastic error injection* trains the network to correct randomly injected spelling mistakes. We recommend best configuration and approach based on evaluation on two training datasets and a sample of real mistakes.

We organize the rest of the paper as follows: In Section II, we survey the state-of-the-art techniques in detecting and correcting spelling mistakes in European, Indo-Iranian and the Arabic languages. In Section III, we review the basic concepts of recurrent neural networks, long short-term memory, and the sequence transcription problem. Section IV details the experimental setup used in this work, the datasets, the used training approaches, and the performance evaluation metrics. Section V presents and discusses the results of our experiments. We provide a summary and conclude the paper in Section VI.

II. RELATED WORK

The natural language processing community has an ongoing interest in spell checking and correction. Traditionally, post-OCR spell check and correction has been a driving force behind research and application. Yet, in the past decade, the proliferation of social media and the high reliance on instant messaging demand more efficient and accurate spell

checkers and on-the-fly accurate correction. The accuracy level of spelling and grammatical mistakes correction varies in maturity between different languages.

It is worth noting that research experiments have been usually evaluated based on artificially-created or proprietary corpora and less so on a corpus of authentic misspellings [15]. Moreover, generic spelling checkers (GSCs), such as the ones packaged in popular text editors like Microsoft Word, are designed for native writers and as such fail to detect and correct mistakes commonly introduced by second language learners [16]. In an ever-interconnected world where hundreds of millions of people are bilingual and multilingual, designing accurate spell-checkers is more challenging. In this section, we review some of the most recent works in spelling correction for three world language groups.

A. Spelling Correction for European Languages

Whereas spelling correction for the English language is well-established compared to other languages, most of this research was evaluated on proprietary corpora of native English texts or well-formed texts with artificially injected errors. Yet, spelling correction of non-native speakers is far more challenging as they feature multi-character edits compared to a single-character edit produced by native speakers [17]. To this end, the authors in [15] developed a minimally supervised model based on contextual and non-contextual features. These features include orthographic similarity, phonetic similarity, word frequency, n-grams and word embeddings among others. Notably, they present and use a corpus of real-world learner essays from the TOEFL exam, and further test their model on out-of-domain medical notes. They report an accuracy level of 88.12% on the TOEFL set, and 87.63% on the medical set. The authors in [18] propose a nested RNN model for English word spelling error correction. They generate pseudo data based on phonetic similarity to train the network. Their proposed system has a precision of 71.77%, a recall rate of 61.26%, and an $F_{0.05}$ score of 69.39%.

D’hondt *et al.* [19] employ many-to-many character sequence learning network using LSTM for French text cor-

rection. Their model stacks two LSTM layers: an encoder layer that reads the sequence of characters, and a decoder layer that generates the output. They further use a drop-out layer to enhance performance. They train and evaluate their system on a dataset of OCRed French medical notes using two models that introduce noise and confusion into the text. They report an accuracy rate of 73% for the former and 71% for the latter model. The same authors extended their work by using BiLSTM [20], and used various corpora based on structured English, structured French, and free-text French with artificially corrupted strings. They report accuracy rates no less than 85% for the structured English and French, and 60% for the free-text French. They show that for an original text with a character error rate (CER) of 34.3%, the BiLSTM system reduces the CER to 7.1%.

B. Spelling Correction for Indo-Iranian and Asian Languages

Dastgheib *et al.* [21] introduced Perspell; a semantic-based spelling correction system for the Persian language which is based on an n-gram model. Perspell handles both real-word and non-word errors. The authors' experiments show that for non-word errors, the precision, recall, and F_1 score are 87.7%, 88.9%, and 88.3%, respectively; while for real-word errors, the authors report a precision of 92.4%, a recall rate of 93%, and an F_1 score of 92.6%.

More recently, Yazdani *et al.* [22] use dictionary-based methods to detect word misspells. They rely on a generic Persian dictionary and a specialized medical dictionary as their system is oriented towards health care applications, specifically ultrasound reports. They employ an n-gram model to dictate suggestions based on orthographic and edit distances. They test their system on actual ultrasound free-text reports and achieve a detection performance of up to 90.29% and a correction accuracy of 88.56%.

Salavati *et al.* [23] introduce Rênûs, a spell checker for the Sorani dialect of the Kurdish language. The error detection phase in Rênûs is based on an n-gram frequency model, and the error correction phase is based on the edit distance, as a measure of similarity and frequency. The authors carried out experiments to investigate error correction once with the use of lexicon and another without. They report a correction accuracy of 96.4% for the former and 87% for the latter case.

The work in [24] introduces the SCMIL system which stands for sequence-to-sequence text correction model for Indic languages. SCMIL uses an attention model with a bidirectional RNN encoder and attention decoder. The decoder is trained end-to-end and it has a character-based representation on both encoder and decoder sides. They have synthesized a dataset from the Hindi and Telugu languages with data lists comprised of a maximum of five words. They subsequently introduced errors which include insertion, deletion, substitution, and word fusion. The authors show that SCMIL has an accuracy rate of 85.4% for the Hindi language and 89.3% for the Telugu language.

Zhang *et al.* [25] proposed a system for Chinese spelling error detection which consists of a network for error detection and a network for error correction based on BERT. The two networks are connected to each other through a technique they called soft-masking. For a training set of five million examples,

the authors' error detection system achieved an accuracy of 80.8%, a precision of 65.5%, a recall of 64% and an F_1 score of 64.8%. The error correction system; however, achieved an accuracy of 77.6%, a precision of 55.8%, a recall of 54.5% and an F_1 score of 55.2%.

C. Spelling Correction for the Arabic Language

Most recent works for Arabic spelling detection and correction still use traditional techniques in the field of natural language processing (NLP). For example, the authors in [26] introduced a spell checker which targets both lexical and semantic spelling mistakes. He uses a sequential combination of approaches including lexicon-based, rule-based, and statistical-based methods. He achieves an F_1 score of 67%.

Al-Shneifi *et al.* [27] developed a cascade system called Arib that detects and corrects a range of spelling errors. Errors that are discovered by Arib include: edit, add, split, merge, punctuation, phonological, and other observed common mistakes. They employed two core models: a probabilistic model based on Bayes probability theory and a Levenshtein distance-based model. They further add three extra models; two of which are based on 3rd party error detection tools: MADAMIRA and Ghaltawi, and the third additional module is a rule-based correcter derived from analyzing samples of the QALB database. Overall, Arib has an F_1 score of 57.8%, and precision and recall rates of 66.6% and 51.1%, respectively.

Mubarak and Darwish [28] also used a cascaded approach for word-level errors, followed by punctuation correction. For word-level correction, the authors used a statistical character-level transformation model and a language model to handle letter insertions, deletions, and substitutions and word merges. The author subsequently use a case-specific system aided by a language model to handle specific error types such as dialectal word substitutions and word splits. For punctuation recovery, the authors employ a simple statistical word-based system and a conditional random fields sequence labeler (CRF). For different experiments, the authors were able to achieve a precision rate up to 71.7%, a recall rate up to 60.32%, and an F-measure up to 63.43%.

Bouamor *et al.* [29] introduced another hybrid system that is based on a morphology-based corrector; rule-based linguistic techniques, language modeling, statistical machine translation (SMT), as well as an error-tolerant finite-state automata method. They target common error types which include split, delete, edit, merge, move and add errors based on the 2014 QALB set. They report an F_1 score of 68.4%.

Noaman *et al.* [30] developed a hybrid system based on the concept of confusion matrix and the noisy channel spelling correction model. They automatically detect and correct Arabic spelling errors of the edit and split types based on the QALB dataset. They report a word error correction rate up to 89.7%.

Zahui *et al.* [31] introduced Al-Mossahih tool that detects and corrects one-letter typographical and phonetic transcription errors. Their detection phase is dictionary-based where a collection of around two million words are sorted in alphabetical order. The correction module encompasses four techniques: one is based on a correspondence table between pairs of commonly confused characters, the second is permutation-based where all possible words from the word letters are

generated, the third is a neighborhood module which considers letters whose keys are nearby on the keyboard, and finally a language model that deals with word locations within a sentence. Al-Mossasih tool has a word error detection rate of 74.75% and a correction rate of 80.2%.

Semantic errors have been addressed by few recent works. A major approach is based on confusion matrices, which despite being powerful, they limit the number of errors that can be detected and corrected. Al-Jefri and Mahmoud [32] compiled a corpus of 7.4 million words from the set of words most confused by non-native Arabic speakers and from the set of mis-recognized words by Arabic OCR systems. The authors compiled these words into 28 confusion sets with assigned probabilities. Errors detection only targets words listed in the confusion matrices and error correction is based on picking the word with highest probability using the computed n-gram model. They report an average accuracy of 95.4%.

For non-confusion set-based approaches, Zribi and Ahmed [33] detected semantic errors through the use of four combined statistical and linguistic methods. They have introduced semantic errors on a set of sentences collected from economic articles from the Egyptian Al-Ahram newspaper. The semantic errors were all a single edit away from the correct word. The reported detection performance was 90% and 83% for precision and recall, respectively.

Rokaya [34] introduced a small variation into the previous method by using the power link method instead of traditional frequency to detect and correct semantic errors coupled with confusion sets as a hybrid approach. They only report results for the detection stage where their system achieves 94.35% and 85.57% for precision and recall, respectively.

More recently, Watson *et al.* [35] utilized sequence-to-sequence models and character and word embeddings for Arabic Text Normalization. Azmi *et al.* [36] combine the language model with machine learning in the detection stage. For the correction step, they only use a language model. They have used word n-grams as features which are subsequently fed into a support vector machine classifier (SVM) to detect and mark words with semantic errors. For the correction step, they generate candidate words which are one-edit distance away from the erroneous word. The candidates are ranked and sorted based on the n-gram language model and then they select the best suggestion accordingly. Their system has an F_1 score of 90.7%, and an precision and recall rates of 83.5% and 99.2%, respectively.

Alkhatib *et al.* [37] recently used an LSTM model to detect and correct spelling and grammatical mistakes at the *word-level*. Their model uses word-embeddings and a polynomial classifier. They report an F_1 score of 93.89%, and for morpho-syntactic mistakes pertaining to word form, noun number, verb form, and verb tense, they report a precision of 95.6% and a recall rate of 94.88%. Solyman *et al.* [38] employed CNNs for the automatic correction of Arabic grammar. After fine-tuning their different developed and tested models, the authors achieved a precision of 80.23%, a recall rate of 63.59%, and an F_1 score of 70.91. Kuznetsov and Urdiales [39] proposed a method of performing spelling correction on short input strings, such as search queries or individual words using denoising auto-encoder transformer model to recover the original query.

They used datasets for four languages and achieved an accuracy of 83.33% (Arabic), 91.83% (Russian), 93.97% (Greek), and 94.48% (Setswana).

III. MACHINE LEARNING AND SEQUENCE TRANSCRIPTION

A general definition of sequence transcription is the process of transforming an input sequence into a corresponding output sequence. Within the context of machine learning spelling detection and correction, the input sequence is the set of letters forming the text that may have spelling errors. The corresponding output sequence is the text on which the machine learning algorithm attempted corrections. Sequence transcription is quite common in similar problems in the fields of language translation, voice recognition and diacritizing Arabic texts [12]. In all these applications, we need to infer relationships and provide outputs depending on past input data. Therefore, the need to preserve correlations between data points in the sequence is necessary.

Recurrent neural networks (RNN) provide the capability to learn from data sequences and consequently infer relevant output data. In this section, and for the sake of completeness, we briefly review RNN and a special RNN network cell called long short-term memory (LSTM) that we adopted in this work. We will further provide details of how we handle and process the input sequence prior to its use in the LSTM network.

A. Basic Recurrent Neural Networks

In general, RNN maintain hidden states that are functions of previous input. This enables such networks to infer outputs from past sequences making them quite suitable for applications where we require sequence transcription. A standard RNN cell can be described by two equations that relate the input sequence $x_t \in (x_1, x_2, \dots, x_T)$ to the hidden states $h_t \in (h_1, h_2, \dots, h_T)$ and output sequence $y_t \in (y_1, y_2, \dots, y_T)$. Equations 1 describe the standard RNN model:

$$h_t = \sigma(W_h x_t + U_h h_{t-1} + b_h) \quad (1a)$$

$$y_t = W_y h_t + b_y \quad (1b)$$

where W , U , and b denote the weight and bias matrices. We can clearly observe that the current hidden state h_t depends on both the current input x_t and the previous hidden state h_{t-1} . The model uses a sigmoid function σ to limit the output within a prefixed range. Hyperbolic tangents \tanh , for example, limit the output to between -1 and 1 , whilst logistic sigmoid limits the output to between 0 and 1 . Both functions are often used in RNN among others. In this paper, we strictly use the σ symbol to denote a logistic sigmoid whilst we express the hyperbolic tangent as \tanh . As stated earlier, the hidden state equips the network model with the ability to learn from input sequences, remember past data and infer outputs.

However, if the input sequence is long, and the output data to be inferred depends on a much older (past) input data sequence (*i.e.*, the gap or the distance in sequence between the inferred output and relevant input data is large), RNN in its standard form might not be able to deliver an accurate desired output sequence. To this end, researchers developed the refined RNN cell which they called long short-term memory

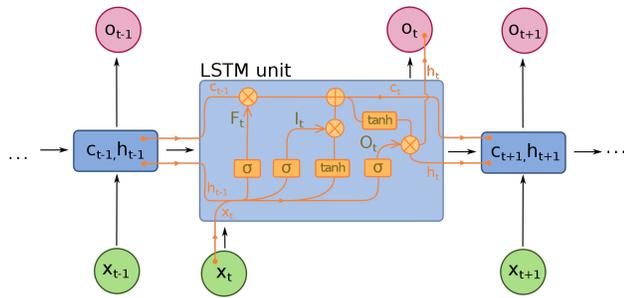


Figure 1. Basic Structure of an LSTM Cell [41]

(LSTM) [40]. LSTM can handle much longer sequences as well as avoid an inherent problem in standard RNN: the vanishing gradient problem. The vanishing gradient problem halts the update of network weights when the gradient is too small; thus stopping the learning stage quite early resulting in a poor network model.

B. Long Short-Term Memory Cells

The internal architecture of an LSTM cell vastly improves on the basic RNN cell by incorporating a *set* of gates which govern the operation of each individual cell. These gates equip the cell with the capacity to work with short or long-term contexts. LSTM cells are relatively insensitive to long gaps or large distance between the output to be inferred and relevant old data in the sequence. Figure 1 illustrates the internal architecture of an LSTM cell. An LSTM cell has an *input gate I*, a *forget gate F*, an *output gate O*, and a *cell activation unit C*.

We provide the governing equations of these gates in Equations 2. For each gate type, we use the small letter notation to denote the output of the gate at time t :

$$i_t = \sigma(W_i x_t + U_i h_{t-1} + V_i c_{t-1} + b_i) \quad (2a)$$

$$f_t = \sigma(W_f x_t + U_f h_{t-1} + V_f c_{t-1} + b_f) \quad (2b)$$

$$c_t = f_t \circ c_{t-1} + i_t \circ \tanh(W_c x_t + U_c h_{t-1} + b_c) \quad (2c)$$

$$o_t = \sigma(W_o x_t + U_o h_{t-1} + V_o c_t + b_o) \quad (2d)$$

$$h_t = o_t \circ \tanh(c_t) \quad (2e)$$

where W , U , V , and b denote the weight and bias matrices and the initial values for $c_0 = h_0 = 0$. The \circ operator denotes the Hadamard (element-wise) product. Similar to the basic RNN cell, an update to a short-term state h_t or long-term state c_t depends on the current input x_t , and previous short-term and long-term memory states h_{t-1} and c_{t-1} .

LSTM is likely to quickly over-fit the training data; thus rendering them less powerful in predicting correct outputs. Dropout is a computationally cheap way that reduces overfitting. It further improves generalization error and model performance in all deep neural networks. Dropout simply

TABLE III. LETTERS CONVERTED TO INTERMEDIATE CODES

Letter(s)	Intermediate Code
اء	J
وا	A
و	O
ت	T
ه	H

implies probabilistically excluding nodes from activation and weight updates while training a network.

The bidirectional LSTM is a version of the LSTM architecture that exploits future contexts as well as past contexts. That is, in sequence transcription problems, it could be beneficial to have access to subsequent (future) sequences to infer correct current outputs. In BiLSTM, we couple the conventional unidirectional LSTM network that works with past sequences with another unidirectional network that works with future sequences. We train the former in a forwards fashion whilst the latter in a backwards fashion. The final output is a concatenation of both the forward and backward LSTM networks.

C. Sequence Processing and Data Encoding

We can classify RNN based on the relationship between the input and output sequence lengths into four types:

- *One-to-one* networks where the length of the input sequence matches the length of the output sequence.
- *Many-to-one* networks where input sequences are transcribed into one final output, for example classification problems.
- *One-to-many* networks where one input vector is used to produce an output sequence.
- *General many-to-many* networks where the number of items in the input sequence differs than that in the output sequence. We use encoder-decoder architectures to tackle these transcription problems by using an intermediary fixed-length vector. The encoder maps a variable-length source sequence to the intermediary fixed-length vector, and the decoder maps the vector representation to a variable-length target sequence.

We can model the Arabic spelling correction problem as a one-to-one problem. Intuitively, this should be straightforward as most soft Arabic spelling errors result from confusing shapes of الهمزات (al-hamzāt) or the similar sounding characters at the end of the word. Simple one-character replacement fixes the misspelled word. However, some of these soft mistakes fall under the addition/omission type. For example, the spelling error in cases 4 and 15 in Table II add or omit the *alef* at the end of the word after a *waw*. Similarly, case one corrects writing *middle hamza* on an *alef* (one letter) by writing it properly following an *alef* (two letters). These particular cases result in differing-length sequences. Ordinarily, the encoder/decoder RNN architecture handles this well, yet with extra cost and overhead.

To mitigate and simplify our approach, we propose a simple yet effective technique that maintains one-to-one sequencing by processing the input sequence stream prior to applying it to the neural network, and then post-processing the output sequence to restore readable Arabic form. We convert some letters and two-letter combinations to intermediate arbitrary codes of English letters, as specified in Table III. The conversion involves letters that are prone to the spelling mistakes under study: the combination *alef-hamza* (ء) and the letters at word ends (و, بو, ت and ه). We emphasize that this conversion is positional in nature; that is; we convert the *waw-alef* combination (وا) to an 'A' only when it appears at word endings where the associated error frequently occurs. Should this combination appear in the middle, no conversion is performed. However, we convert the letter combination *alef-hamza* to 'J' wherever it occurs in the word. This combination is susceptible to soft mistakes both in the middle and at the end of the word.

In our machine training approach, we use this converted sequence as the target sequence and a copy of it as the input sequence after injecting some artificial spelling mistakes (refer to Section IV-B). These sequences are stored using the Unicode UTF-8 encoding. However, when presented to the neural network, they are put in 3D ($B \times T \times C$) tensors (dense matrices). Each tensor holds a batch of B sequences of a maximum length $T = 400$ characters. Note that we wrap sequences longer than 400 characters to improve training performance. The third dimension encodes each of the C distinct characters using one-hot encoding.

IV. EXPERIMENTAL SETUP

In this work, we use an experimental setup similar to the one used in our past research [42]. We list the specifications of the experimental platform in Table IV.

A. Machine Learning Model Configurations

We develop our machine learning models using Python deep learning libraries ensuring we use the latest versions of the algorithms. Specifically, we use Keras with TensorFlow at the backend. Given that we tackle the problem of Arabic spelling correction at the character level, we adopt BiLSTM RNN. These networks can handle longer sequences which can be beneficial for correcting misspelled words based on past and future context. We develop and compare three models. The baseline model has only two BiLSTM layers. We add an input masking layer before the two hidden BiLSTM layers, and connect their output to a fully-connected (dense) output layer, as shown in Fig. 2.

The second model maintains the same settings of the previous one but further employs the *dropout* method to reduce or avoid the over-fitting problem. The third and final model has four BiLSTM hidden layers instead of two and also uses *dropout*. We use the same configuration settings for the three models. Each bidirectional layer has 256 cells. We use the *softmax* as the activation function of the output layer and the *RMSprop* optimizer in training. We use categorical cross entropy as the loss function, and a batch size of 64 sequences with a sequence length of 400 while wrapping longer sequences similar to our work in [12]. To combine the forward and backward layers in our BiLSTM layers, we use

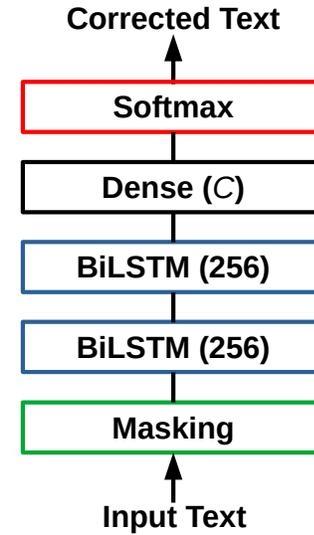


Figure 2. The Base 2-Layer Network Model

concatenation. For the models that use dropout, we use the base settings of dropout = 0.1 and recurrent_dropout = 0.3. We set the training time to a maximum of 50 epochs with early stopping and 5-epoch patience. We show the skeleton code of the 2-Layer model with dropout in Listing 1.

B. Data Sets

We use two widely used datasets for Arabic NLP [43], [44], [45], [46], [47], [48], [49] to train, validate, and test our proposed BiLSTM models. The first is a processed subset of the Tashkeela corpus as extracted in [50]. The second is from the Linguistic Data Consortium's (LDC) Arabic Treebank (LDC2010T08), specifically speaking: the Arabic Treebank Part 3 (ATB3) v3.2 [51]. We will now-forth refer to this dataset as ATB3 in this paper. The major difference between the Tashkeela and ATB3 datasets is the form of the Arabic text they consist of. Tashkeela mainly contains 55K sequences from classical Arabic texts. On the other hand, ATB3 contains samples of modern standard Arabic of 599 distinct news-wire stories from the Lebanese publication An-Nahar. We show in Table V the characteristics of the two datasets in terms of sequence count, word count, character count, average words per sequence, average letters per word, and fraction of sequences shorter than or equal to 400 characters.

In Table VI, we present the Arabic letters and their variations that account for the majority of the soft spelling errors that we presented in detail in Section I and cited in examples in Table II. We break them down in order of their absolute frequency within the dataset texts relative to the character count. We also break them down in terms of their relative appearance to each other within the same texts. This table demonstrates that about one fifth of the characters are involved in the common soft spelling mistakes under study and that the relative frequencies of these letters are highly skewed ranging from 0.13% to 58.23%.

We split the ATB3 dataset as proposed by Zitouni *et*

TABLE IV. EXPERIMENTAL PLATFORM SPECIFICATIONS

Aspect	Specification
CPU	Intel Core i7-9700KF @ 3.6 GHz, 8 cores, 12 MB cache
GPU	Nvidia GeForce RTX 2080 @ 2.1 GHz, 2944 CUDA cores, 8 GB memory
Memory	32 GB DDR4-SDRAM @ 2666 MHz
OS	Ubuntu 20.04 LTS, 64-bit
Libraries	Python 3.8.2, TensorFlow 2.2.0, Keras 2.3.0-tf

Listing 1: A 2-Layer BiLSTM Model with Dropout

```
1 model = Sequential()
2 model.add(Masking(mask_value = 0, input_shape=(seq_len, num_inp_tokens)))
3 model.add(Bidirectional(LSTM(256, return_sequences=True, dropout=0.1, recurrent_dropout=0.3),
4     merge_mode='concat'))
5 model.add(Bidirectional(LSTM(256, return_sequences=True, dropout=0.1, recurrent_dropout=0.3),
6     merge_mode='concat'))
7 model.add(TimeDistributed(Dense(num_tar_tokens, activation='softmax')))
8 model.compile(loss='categorical_crossentropy', optimizer='rmsprop', metrics=['acc'])
```

al. [52] such that we use the first 509 news-wire stories, in *chronological* order, to train the model and use the remaining 90 stories to validate and test the model. This accounts for 22,170 sequences for training and 3,857 sequences for validation. Similarly, we split the Tashkeela dataset into 50K lines for training, 5,000 lines for testing. In our previous work [12], we analyzed the best maximum sequence length to use based on the same datasets. We found that a maximum sequence of 400 characters provides the best speed versus accuracy trade off. We consequently adopt this sequence length in this work as well.

In addition to the above datasets, we test the proposed solutions using samples of real soft spelling mistakes (Test200). These samples were collected in a previous work [8] and are summarized in Table VII. They have a challenging collection of soft spelling mistakes with an average of 6.5 mistakes per sequence.

C. Training Approaches

We have experimented with the following two approaches to train BiLSTM networks to correct the soft spelling mistakes.

- 1) **Transformed input:** This approach trains the BiLSTM network to predict the correct form of the letters under study given unified transformed input. Once the sequences are converted to the intermediary form, as described in Section III-C, we transform the letters that are often confused with each other into one final

form. We show the used mappings of the letters affected by this transformation in Table VIII. All *hamza* forms are transformed to plain *hamza* (ء), the *heh* and *teh* forms at word ends are transformed to *teh marbuta* (ة), *waw-alef* at word ends are transformed to *waw* (و), and *alef* at word ends are transformed to *alef maksura* (ى).

- 2) **Stochastic error injection:** This approach trains the network to correct artificial errors randomly injected in the input sequences. We inject errors in the input sequence by replacing the target letters pertaining to the soft Arabic spelling mistakes. With an *error injection rate* p , a letter belonging to the four groups shown in Table VIII is randomly replaced by one of the letters in its group. For example, we replace *alef maksura* with an *alef* and vice versa (cases 2 and 3 in Table II). We have investigated using multiple error injection rates p as described in Section V. For example, 10% of the letters under investigation are replaced with $p = 10\%$.

D. Evaluation Metrics

We measure and evaluate the performance of BiLSTM networks using their computational time and multiple performance metrics, namely: *accuracy*, *precision*, *recall*, F_1 *score*, character error rate (*CER*), and word error rate (*WER*). These metrics are quite common [36], [50], [53], [54], [37] in evaluating the performance of solutions pertaining to error detection and correction, voice recognition, and similar sequence-based problems.

The first four measures are readily computed and understood through means of a simple confusion matrix that we will explain within the context of our work. At the character level, any character in any sequence can either be correctly spelled or not. When comparing the predicted outcome of our model with the actual (target) sequence, we can thus have four cases that we show in Fig. 3. For a certain character c , the counts of these four cases are:

TABLE V. TASHKEELA AND ATB3 DATASETS CHARACTERISTICS

Criterion	Tashkeela	ATB3
Sequence count	55K	26K
Word count	2,312K	305K
Character count	12,464K	1,660K
Words per sequence	42.1	11.3
Letters per word	4.0	4.6
Sequences \leq 400 chars.	84.1%	99.9%

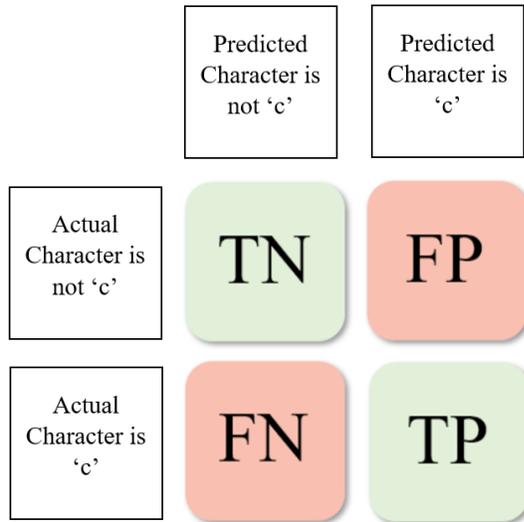


Figure 3. Confusion Matrix of Correct and Erroneous Characters between Predicted and Actual Sequences

- True positive (TP): The number of times character c is correctly predicted as c .
- True negative (TN): The number of times characters other than c are predicted correctly.
- False positive (FP): The number of times characters other than c are incorrectly predicted as c .
- False negative (FN): The number of times character c is incorrectly predicted as another character.

It is evident that we need our model to maximize the correct cases TP and TN (cases shown in green). We desire our model also to keep the incorrectly predicted characters FP and FN to a minimum (cases shown in red). This readily translates to the

TABLE VI. THE SUBSET OF TARGET ARABIC LETTERS (AND THEIR VARIATIONS) UNDER STUDY AND THEIR ABSOLUTE AND RELATIVE FREQUENCIES WITHIN THE DATASETS

Letter(s)	Frequency		Relative Frequency	
	Tashkeela	ATB3	Tashkeela	ATB3
ا	8.38%	11.55%	46.96%	58.23%
ه	2.55%	0.58%	14.31%	2.93%
أ	2.41%	1.70%	13.52%	8.57%
ة	1.22%	2.49%	6.81%	12.55%
ل	0.93%	0.79%	5.21%	3.97%
ح	0.68%	0.24%	3.79%	1.21%
ى	0.67%	0.72%	3.77%	3.65%
ت	0.40%	0.81%	2.24%	4.11%
اء	0.18%	0.24%	1.03%	1.24%
نج	0.18%	0.40%	1.03%	2.04%
آ	0.07%	0.08%	0.40%	0.41%
ء	0.07%	0.03%	0.37%	0.13%
ؤ	0.06%	0.14%	0.37%	0.69%
وا	0.04%	0.06%	0.34%	0.28%
Subtotal	17.85%	19.83%	100.00%	100.00%
Other chars.	82.15%	80.17%	0.00%	0.00%
Total	100.00%	100.00%	100.00%	100.00%

definition of the accuracy metric that we present in Equation 3:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (3)$$

The precision metric is the ratio of correct predictions of character c to total number of characters predicted as c and is given by Equation 4:

$$Precision (P) = \frac{TP}{TP + FP} \quad (4)$$

The recall metric is the ratio of correct predictions of character c to the total count of actual c characters. Equation 5 mathematically defines the recall as:

$$Recall (R) = \frac{TP}{TP + FN} \quad (5)$$

The F_1 score combines the precision and recall metrics into one score by applying the weighted harmonic mean on both giving equal weights to each. We get the F_1 score using Equation 6:

$$F_1 = 2 \times \frac{P \times R}{P + R} \quad (6)$$

We additionally use a new metric (FP/Changes) to assess the prediction error rate with respect to the number of replacements in the input sequence. We divide the number of false positive cases of a letter over the number of times this letter was changed in the input sequence. The accuracy is reported in this work for the entire set of characters. Whereas, precision, recall, F_1 , and FP/Changes scores are reported as weighted averages of the target letters shown in Table VI.

We also evaluate our models using the character error rate (CER) and word error rate (WER). CER is the percentage of letters that are misspelled whereas WER is the percentage of misspelled words. A word is considered misspelled if it has at least one incorrectly spelled letter.

TABLE VII. TEST200 TEST SET OF REAL SOFT SPELLING MISTAKES

Criterion	Count
Sequence count	200
Word count	2,443
Character count	24,002
Number of mistakes	1,306
Mistakes per sequence	6.5

TABLE VIII. LETTERS CHANGED IN THE “TRANSFORMED INPUT” APPROACH

Intermediate Forms	Mapped To
[ء, آ, أ, ؤ, ة, ا, and J (اء)]	ء
[ة, T (ت), and H (ه)]	ة
[O (و) and A (وا)]	O (و)
[ا and ى]	ى

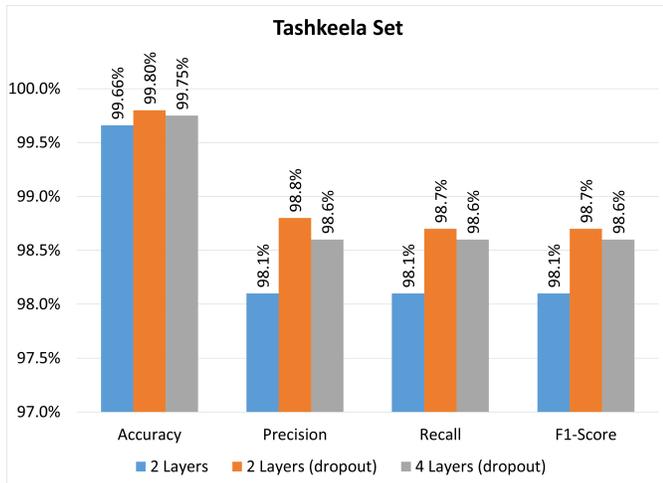


Figure 4. Performance of Three BiLSTM Networks using the Transformed Input Training on the “Tashkeela” Set.

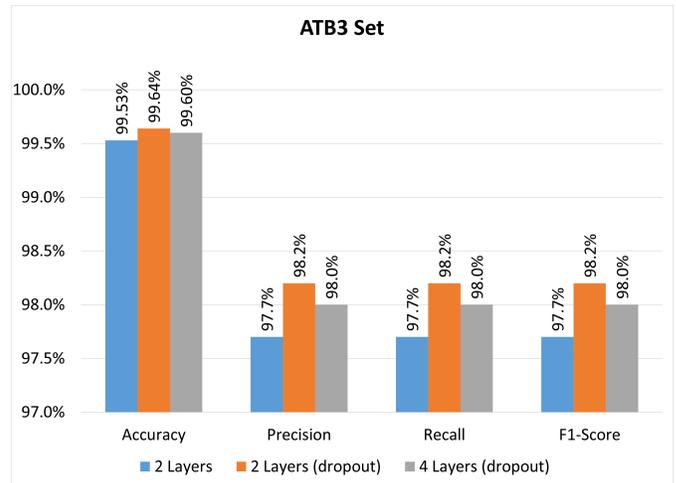


Figure 5. Performance of Three BiLSTM Networks using the Transformed Input Training on the “ATB3” set.

V. RESULTS AND DISCUSSION

This section presents the results of evaluating alternative network topologies and the two training approaches proposed above. This section also presents and discusses the detailed performance evaluation results.

A. Network Topology

We started by evaluating three network topologies: the base network of two BiLSTM layers, another that adds the dropout technique to the base network, and a third that extends the second network to having four BiLSTM layers instead of two. These three alternatives are inspired by experience from previous work, e.g., [42]. In this initial evaluation, we use the *transformed input* training approach. For each of these three networks, we use the Tashkeela and ATB3 sets to separately train two separate instances of the BiLSTM network; one trained for classical Arabic, and another for MSA Arabic. We repeat this procedure for the three networks using the same training sequences and compare their performance in terms of *accuracy*, *precision*, *recall*, and F_1 score on the test set.

In Fig. 4, we show these metrics for the networks trained with the Tashkeela set. Despite the results being marginally close, we concur that networks with the *dropout* technique employed perform generally better than the one without. We observe that there is no large difference between the performance of the 2-layer and 4-layer networks with dropout, with only a slight edge towards the 2-layer network. Fig. 5 similarly illustrates the same metrics but for the case when we evaluate the BiLSTM networks using the ATB3 set. We notice that our observations and conclusions for the Tashkeela set carry over to the ATB3 set. However, the accuracy with the ATB3 set is generally lower than the accuracy on the larger Tashkeela set.

We notice also that the accuracy is higher than the other three metrics because it is calculated for the entire character set while the other metrics are weighted averages of the target letters only. For the remaining experiments, and given the

slight performance edge and lesser complexity of the 2-layer BiLSTM network with dropout used, we conduct and evaluate the remaining experiments and report their results based on this model only.

B. Training Approach

In this section, we present the performance of the adopted 2-layer network on the two proposed training approaches: *transformed input* and *stochastic error injection*. The error injection rate p correlates with the network’s ability to correct errors. Therefore, we experimented with multiple rates, e.g., 2.5%, 10%, and 40%, alongside the transformed input approach.

For this evaluation, we show the results also in terms of *accuracy*, *precision*, *recall*, and F_1 score. Recall that these metrics are found not only from the letters that we actually changed, but also include the whole letters in the subset under study. Therefore, we expect that the inclusion of all the letters that could possibly be changed and those actually changed dilutes the results. For example, for the case when the error injection rate is 2.5%, we already know that a high percentage of the remaining 97.5% letters are correct and match the target output.

Despite the expected result dilution, we have to present these results for in most cases the analysis and the interest is in the overall output character sequence and that it should be error free. Given these disproportional ratios between unchanged letters and error-injected letters, we expect better performance with lower error injection rates. We indeed observe these results in Figures 6 and 7. The best performance appears here for the network with the smallest error injection rate of $p = 2.5\%$.

It is difficult to make solid conclusions about the two training approaches from the previous data alone. So we compare the two training approaches using the *FP/Changes* ratio. This allows us to observe more insights that we could not deduce from the diluted results in Fig. 6 and 7. We present

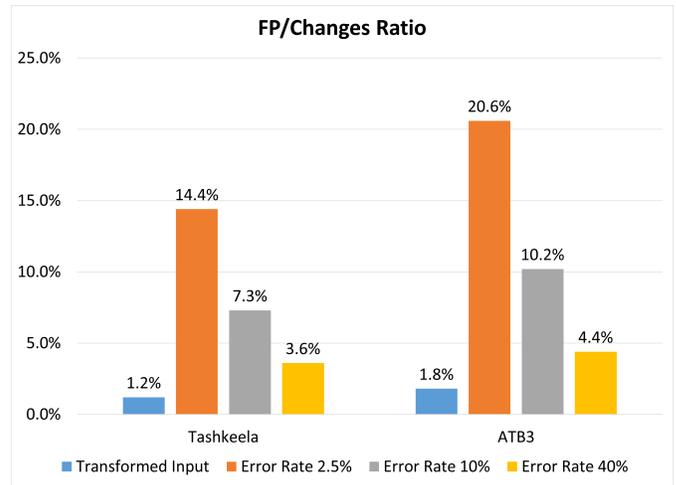
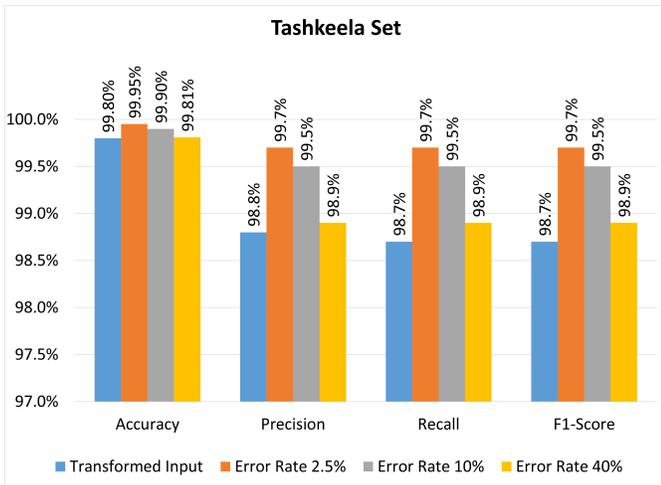


Figure 6. Performance of the Transformed Input and Stochastic Error Injection Training Approaches on the “Tashkeela” Set using the 2-Layer with Dropout Network. Three Error Injection Rates are Investigated.

Figure 8. The *FP/Changes* Ratio of the Two Training Approaches on the Tashkeela and ATB3 Sets.

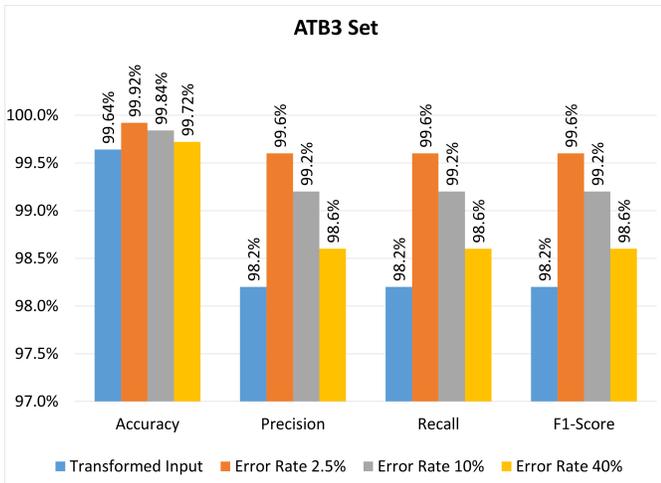


Figure 7. Performance of the Transformed Input and Stochastic Error Injection Training Approaches on the “ATB3” Set using the 2-Layer with Dropout Network. Three Error Injection Rates are Investigated.

this ratio in Fig. 8 for the Tashkeela and ATB3 sets. In this analysis, we easily see that increasing the error injection rate helps train the network to better detect and correct errors. For example for Tashkeela, we observe a decrease of the percentage of false positives from 14.4% to 3.6% as we increase the error injection rate from 2.5% to 40%. Despite the more errors introduced, the network was able to handle them well. In the case of transformed input, we even observe better performance. This is because transforming the most confused characters into one form is another way of introducing different errors into the network training.

In Fig. 9 and 10, we show the character and word error rates for the proposed training approaches. These two metrics are global metrics similar to the accuracy. Therefore, they are generally low because most characters in the input are correct and only need to be passed to the model output as is. We present these metric as any application will handle entire sequences with possible spelling mistakes and attempt to provide an error-free version. Essentially, CER and WER are related so we expect a similar pattern for both. We expect this pattern to resemble that for the analysis presented in above, specifically, $1 - accuracy$. This is due to the ratio of already correct and unchanged characters that are in the input, predicted, and target sequences. Despite this, we observe that for both experiments on the Tashkeela and ATB3 datasets that in the worst case the CER did not exceed 0.36% while the WER did not exceed 1.88%. Referring to the analysis of the datasets that we show in Table V, we note that the average number of letters per word is 4.0 and 4.6 for the Tashkeela and ATB3 sets, respectively. The results we show for WER in Fig. 10 are approximately five times than those for CER in Fig. 9. This is in line with the letters per word statistic for these datasets under consideration.

C. Test200 Results

The above results are not conclusive about which training approach is best. Therefore, we used the BiLSTM models trained on the two training approaches to correct the mistakes in the Test200 set. We use the networks which we trained using the Tashkeela and ATB3 sets here. We report the CER of the predicted Test200 sequences on the eight training configurations shown in Fig. 11. Except for the 2.5% error injection rate, the two training sets provide results within 0.2% for each other. Yet, we note that despite that the reported CER for this external set is quite low, it is an order of magnitude higher than that reported for the test sequences of the Tashkeela and ATB3 datasets. This is because the BiLSTM models were trained using artificially injected errors that might not necessarily always correspond to errors committed by real-users. This

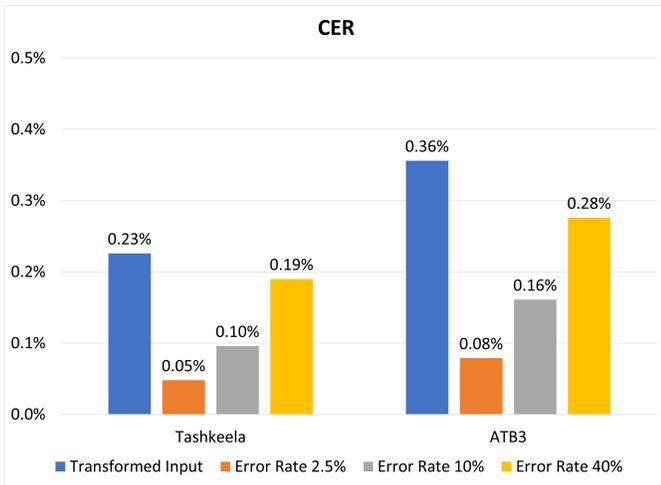


Figure 9. Character Error Rate of the Two Training Approaches on the Tashkeela and ATB3 Sets.

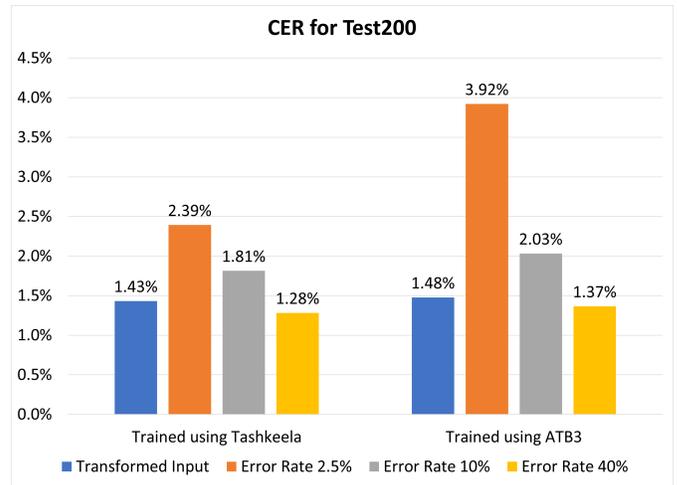


Figure 11. CER of the Two Training Approaches on the Test200 Set for Models Trained on the Tashkeela and ATB3 Sets.

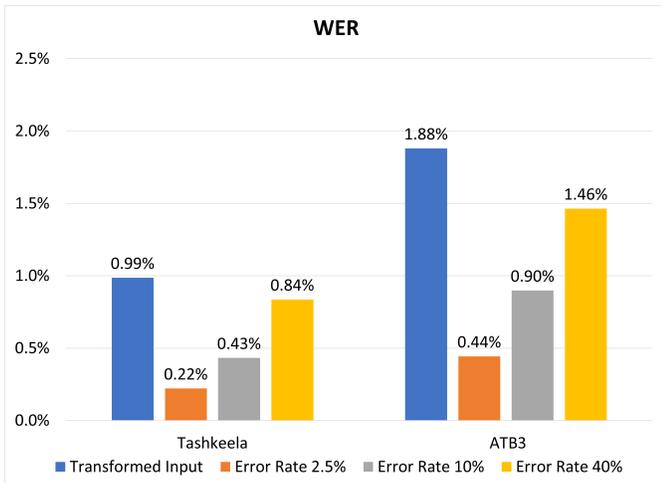


Figure 10. Word Error Rate of the Two Training Approaches on the Tashkeela and ATB3 Sets.

stresses the need for a large Arabic corpora collected from real users, annotated, and corrected by linguistic experts to enrich the Arabic NLP research domain.

Fig. 11 shows also that training on the larger Tashkeela set gives better results and the transformed input and error injection with 40% rate are better than lower rates. The best results with only 1.28% CER is for the case when training using Tashkeela set and 40% error injection rate. To test whether increasing the error injection rate beyond 40% would further decrease CER, we experimented with higher rates (50%, 70%, and 100%). Fig. 12 shows that the model with $p = 40\%$ performs best and yielded the least character error rate. Therefore, we recommend using error injection training approach with $p = 40\%$. We further analyze the results of this approach in the following subsection.

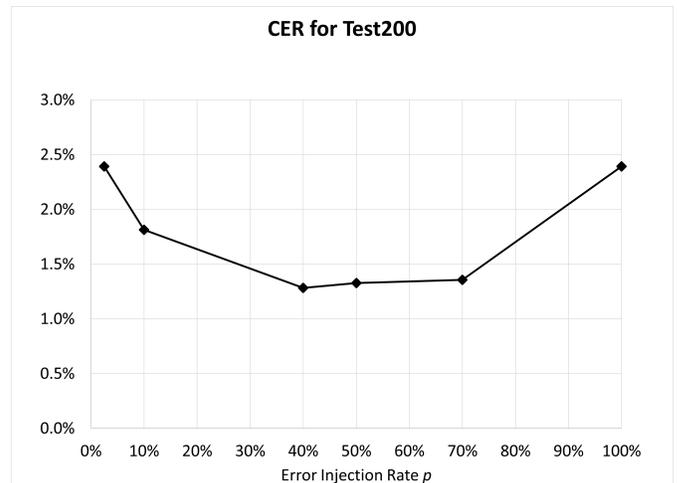


Figure 12. CER of Test200 Set for Models Trained with the Stochastic Error Injection Approach and the Tashkeela Set as a Function of the Error Injection Rate p .

D. Confusion Matrix Results

Fig. 13 shows the confusion matrix for the predicted against actual of the Tashkeela test set of the letters under investigation. We present here this matrix for the best trained model (error injection with $p = 40\%$) and Tashkeela set because it is the largest set we use. This matrix allows us to analyze the letters that our machine learning model mostly confuses with each other. While the majority of the letters are correctly classified and corrected (diagonal), most confusion occurs within the three terminal letters: *heh* (هـ), *teh* (تـ), and *teh marbuta* (ةـ), and *alef with hamza above* (أ) and *alef with hamza below* (إ).

		Predicted Letter													
		وا	ه	اء	و	ت	ء	آ	أ	ؤ	إ	ئ	ا	ة	ى
Actual Letter	وا	264	0	0	3	0	0	0	0	0	0	0	0	0	0
	ه	0	14,699	0	0	66	0	0	0	0	0	0	0	102	0
	اء	0	0	1,083	0	0	0	0	3	0	0	2	7	0	0
	و	9	0	0	3,888	0	0	0	0	0	0	0	0	0	0
	ت	0	123	0	0	2,242	0	0	0	0	0	0	0	30	0
	ء	0	0	0	0	0	381	0	0	0	0	0	0	0	0
	آ	0	0	0	0	0	0	359	25	0	2	0	2	0	0
	أ	0	0	3	0	0	1	20	13,660	19	79	6	31	0	0
	ؤ	0	0	0	0	0	0	0	15	315	0	17	4	0	0
	إ	0	0	0	0	0	0	6	75	0	5,358	1	3	0	0
	ئ	0	0	2	0	0	2	0	4	5	0	957	4	0	0
	ا	0	0	2	0	0	1	6	63	1	3	3	48,505	0	40
	ة	0	151	0	0	46	0	0	0	0	0	0	0	6,881	0
	ى	0	0	0	0	0	0	0	0	0	0	0	107	0	3,735

Figure 13. Confusion Matrix for the 40% Error Injection Training Approach on Tashkeela Test Set.

We list Examples 5, 12, and 16 of these common mistakes in Table II. For example, out of $(151 + 46 + 6,881 = 7,078)$ *teh marbuta* letters, 151 and 46 are wrongly predicted as *heh* and *teh*, respectively.

E. Model Timing

Finally, we report the timing metrics (training time and number of training epochs) of six selected configurations in Fig. 14 and 15 for Tashkeela and ATB3, respectively. In all cases, the training time for the Tashkeela set is between two to five times more than that for the ATB3. It took the longest to train the 4-layer network for both sets; 48.5 hours and 9.2 hours for Tashkeela and ATB3 using the transformed input approach, respectively. When training the network using variable error injection rate, we observe that the training time *generally* increases as this rate increases.

In these experiments, we set the maximum number of training epochs to 100. Yet, all models converged before half the set number of epochs. In contrast to the number of hours parameter, we observe that in some cases it took from $\frac{2}{3}$ to twice the number of epochs to train the same model for the two datasets.

VI. CONCLUSION

In this work, we addressed the problem of correcting common Arabic soft spelling errors. We developed variant configurations of bidirectional LSTM networks with either two or four hidden layers, while using or forgoing the dropout technique. We use sequences of Arabic texts that are either written in classical Arabic (Tashkeela set), or MSA Arabic

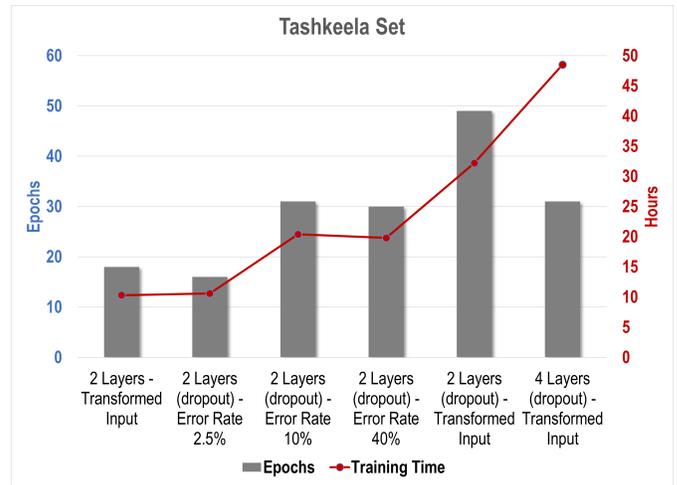


Figure 14. BiLSTM Timing Metrics under different Configurations for the Tashkeela Set.

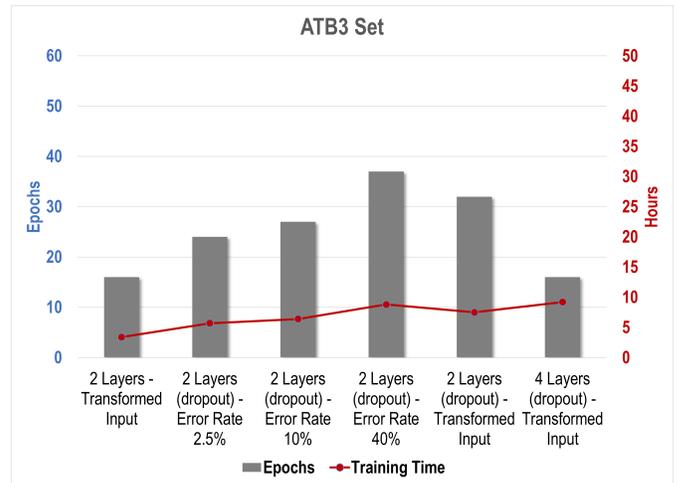


Figure 15. BiLSTM Timing Metrics under different Configurations for the ATB3 Set.

(ATB3 set) to train and validate our models. The 2-layer network with dropout had the highest F_1 score of 98.7% on Tashkeela test set using the transformed input training approach.

We also experimented with a second training approach where we introduce stochastic errors and train the BiLSTM network to correct them. We deliberately varied the percentage of injected errors in the input sequence to assess the BiLSTM network performance and sensitivity given how well it can learn from lower or higher injected error rates. Networks trained with higher error rates (from 2.5%, 10%, through 40%) have worse accuracy, precision, recall, F_1 score, CER, and WER. However, they are more capable of correcting errors as reflected by their *FP/Changes* ratio. Noting that injecting more errors in the input stream beyond 40% does not improve the

network performance in correcting soft spelling mistakes.

The transformed input training approach has better *FP/Changes* ratio than that of the stochastic error injection approach. However, the latter approach is better in correcting the soft spelling mistakes in the Test200 test set with error injection rate of $p = 40\%$. The best result on this test set is an CER of 1.28%. This lowest error rate is for the network trained on Tashkeela set and $p = 40\%$.

For future work, we consider expanding beyond the class of soft Arabic spelling errors. Further, a much larger annotated and pre-processed dataset will open doors to improving accuracy. We could possibly handle Arabic spelling correction and letter diacritization in one problem space.

REFERENCES

- [1] UNESCO, "World Arabic language day," Dec 2019. [Online]. Available: <https://en.unesco.org/commemorations/worldArabicLanguageDay>
- [2] Ethnologue, "Arabic language statistics," 2020. [Online]. Available: <https://www.ethnologue.com/language/ara>
- [3] K. Darwish and W. Magdy, "Arabic information retrieval," *Foundations and Trends in Information Retrieval*, vol. 7, no. 4, pp. 239–342, 2014.
- [4] J. Clement, "Internet: Most common languages online 2019," Jul 2019. [Online]. Available: <https://www.statista.com/statistics/262946/share-of-the-most-common-languages-on-the-internet/>
- [5] T. Tinsley and K. Board, "Languages for the future," British Council, Report ISBN 978-0-86355-883-2, 2017.
- [6] Glosbe, "Transliteration and Romanization utilities," 2020. [Online]. Available: <https://glosbe.com/transliteration/Arabic-Latin>
- [7] K. C. Ryding, "Learning Arabic as a foreign language," 2020. [Online]. Available: <https://Arabic.georgetown.edu/afl/>
- [8] G. Abandah, M. Khedher, W. Anati, A. Zghoul, S. Ababneh, and M. S. Hattab, "The Arabic language status in the Jordanian social networking and mobile phone communications," in *7th Int'l Conf. on Information Technology (ICIT 2015)*, 2015, pp. 449–456.
- [9] "The school of language studies - language categories," 2017. [Online]. Available: <https://2009-2017.state.gov/m/fsi/sls/orgoverview/languages/index.htm>
- [10] G. A. Abandah, A. Graves, B. Al-Shagoor, A. Arabiyat, F. Jamour, and M. Al-Tae, "Automatic diacritization of Arabic text using recurrent neural networks," *International Journal on Document Analysis and Recognition (IJ DAR)*, vol. 18, no. 2, pp. 183–197, 2015.
- [11] S. Alqudah, G. Abandah, and A. Arabiyat, "Investigating hybrid approaches for Arabic text diacritization with recurrent neural networks," in *2017 IEEE Jordan Conf. on Applied Electrical Engineering and Computing Technologies (AEECT)*, Oct 2017, pp. 1–6.
- [12] G. Abandah and A. Abdel-Karim, "Accurate and fast recurrent neural network solution for the automatic diacritization of Arabic text," *Jordanian Journal of Computers and Information Technology*, vol. 6, no. 2, pp. 103–121, 2020.
- [13] M. Attia, P. Pecina, Y. Samih, K. Shaalan, and J. Van Genabith, "Arabic spelling error detection and correction," *Natural Language Engineering*, vol. 22, no. 5, pp. 751–773, Sep. 2016.
- [14] A.-M. H. Al-Ameri, "Common spelling mistakes among students of teacher education institutes," *The Islamic College University Journal*, vol. 1, no. 33, pp. 445–474, 2015.
- [15] M. Flor, M. Fried, and A. Rozovskaya, "A benchmark corpus of English misspellings and a minimally-supervised model for spelling correction," in *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*. Florence, Italy: Association for Computational Linguistics, 2019, pp. 76–86.
- [16] M. Blazquez and C. Fan, "The efficacy of spell check packages specifically designed for second language learners of Spanish," *Pertanika Journal of Social Sciences & Humanities*, vol. 27, no. 2, 2019.
- [17] M. Flor, Y. Futagi, M. Lopez, and M. Mulholland, "Patterns of misspellings in L2 and L1 English: a view from the ETS spelling corpus," *Bergen Language and Linguistics Studies*, vol. 6, 2015.
- [18] H. Li, Y. Wang, X. Liu, Z. Sheng, and S. Wei, "Spelling error correction using a nested RNN model and pseudo training data," *arXiv:1811.00238 [cs]*, Nov. 2018.
- [19] E. D'hondt, C. Grouin, and B. Grau, "Low-resource OCR error detection and correction in French clinical texts," in *Proceedings of the Seventh International Workshop on Health Text Mining and Information Analysis*. Auxtin, TX: Association for Computational Linguistics, 2016, pp. 61–68.
- [20] E. D'hondt, C. Grouin, and B. Grau, "Generating a training corpus for OCR post-correction using encoder-decoder model," in *Proceedings of the Eighth International Joint Conf. on Natural Language Processing*, 2017, pp. 1006–1014.
- [21] M. B. Dastgheib, S. M. Fakhrahmad, and M. Z. Jahromi, "Perspell: A New Persian Semantic-based Spelling Correction System," *Digital Scholarship in the Humanities*, vol. 32, no. 3, pp. 543–553, 03 2016.
- [22] A. Yazdani, M. Ghazisaeei, N. Ahmadinejad, M. Giti, H. Amjadi, and A. Nahvijou, "Automated misspelling detection and correction in Persian clinical text," *Journal of Digital Imaging*, pp. 1–8, 2019.
- [23] S. Salavati and S. Ahmadi, "Building a lemmatizer and a spell-checker for Sorani Kurdish," *CoRR*, vol. abs/1809.10763, 2018. [Online]. Available: <http://arxiv.org/abs/1809.10763>
- [24] P. Etoori, M. Chinnakotla, and R. Mamidi, "Automatic spelling correction for resource-scarce languages using deep learning," in *Proceedings of ACL 2018, Student Research Workshop*. Melbourne, Australia: Association for Computational Linguistics, 2018, pp. 146–152.
- [25] S. Zhang, H. Huang, J. Liu, and H. Li, "Spelling error correction with soft-masked bert," *arXiv preprint arXiv:2005.07421*, 2020.
- [26] M. Mars, "Toward a robust spell checker for Arabic text," in *International Conf. on Computational Science and Its Applications*. Springer, 2016, pp. 312–322.
- [27] N. AlShenaifi, R. AlNefie, M. Al-Yahya, and H. Al-Khalifa, "ARIB@ QALB-2015 shared task: a hybrid cascade model for Arabic spelling error detection and correction," in *Proceedings of the Second Workshop on Arabic Natural Language Processing*, 2015, pp. 127–132.
- [28] H. Mubarak and K. Darwish, "Automatic correction of arabic text: A cascaded approach," in *Proceedings of the EMNLP 2014 Workshop on Arabic Natural Language Processing (ANLP)*, 2014, pp. 132–136.
- [29] H. Bouamor, H. Sajjad, N. Durrani, and K. Oflazer, "Qcnuq@ qalb-2015 shared task: Combining character level mt and error-tolerant finite-state recognition for Arabic spelling correction," in *Proceedings of the Second Workshop on Arabic Natural Language Processing*, 2015, pp. 144–149.
- [30] H. M. Noaman, S. S. Sarhan, and M. Rashwan, "Automatic Arabic spelling errors detection and correction based on confusion matrix-noisy channel hybrid system," *Egypt Comput Sci J*, vol. 40, no. 2, pp. 54–64, 2016.
- [31] A. Zahui, W. Elhor, and M. A. Cheragui, "EL-Mossahih V1.0: A hybrid approach for detection and correction of typographical and phonetic transcription errors in Arabic texts," in *2017 8th International Conf. on Information Technology (ICIT)*. IEEE, 2017, pp. 774–779.
- [32] M. M. Al-Jefri and S. A. Mahmoud, "Context-sensitive Arabic spell checker using context words and n-gram language models," in *2013 Taibah University International Conf. on Advances in Information Technology for the Holy Quran and Its Sciences*. IEEE, 2013, pp. 258–263.
- [33] C. Zribi and M. Ahmed, "Detection of semantic errors in Arabic texts," *Artificial Intelligence*, vol. 195, pp. 249–264, 2013.
- [34] M. Rokaya, "Arabic semantic spell checking based on power links," *International Information Institute (Tokyo). Information*, vol. 18, no. 11, p. 4749, 2015.
- [35] D. Watson, N. Zalmout, and N. Habash, "Utilizing character and word embeddings for text normalization with sequence-to-sequence models," *arXiv preprint arXiv:1809.01534*, 2018.
- [36] A. M. Azmi, M. N. Almutery, and H. A. Aboalsamh, "Real-word errors in Arabic texts: A better algorithm for detection and correction," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 8, pp. 1308–1320, 2019.

- [37] M. Alkhatib, A. A. Monem, and K. Shaalan, "Deep learning for Arabic error detection and correction," *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, vol. 19, no. 5, pp. 1–13, 2020.
- [38] A. Solymán, W. Zhenyu, T. Qian, A. A. M. Elhag, M. Toseef, and Z. Aleibeid, "Synthetic data with neural machine translation for automatic correction in arabic grammar," *Egyptian Informatics Journal*, 2020.
- [39] A. Kuznetsov and H. Urdiales, "Spelling correction with denoising transformer," *arXiv preprint arXiv:2105.05977*, 2021.
- [40] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.
- [41] Wikimedia_Commons, "Long short-term memory," 2007. [Online]. Available: https://en.wikipedia.org/wiki/File:Long_Short-Term_Memory.svg
- [42] G. A. Abandah, M. Z. Khedher, M. R. Abdel-Majeed, H. M. Mansour, S. F. Hulliel, and L. M. Bisharat, "Classifying and diacritizing Arabic poems using deep recurrent neural networks," *Journal of King Saud University - Computer and Information Sciences*, 2020.
- [43] I. H. Ali, Z. Mnasri, and Z. Lachiri, "Dnn-based grapheme-to-phoneme conversion for arabic text-to-speech synthesis," *International Journal of Speech Technology*, vol. 23, no. 3, pp. 569–584, 2020.
- [44] N. Habash, A. Shahrour, and M. Al-Khalil, "Exploiting arabic diacritization for high quality automatic annotation," in *Proceedings of the Tenth International Conf. on Language Resources and Evaluation (LREC'16)*, 2016, pp. 4298–4304.
- [45] I. Bounhas, N. Soudani, and Y. Slimani, "Building a morpho-semantic knowledge graph for arabic information retrieval," *Information Processing & Management*, vol. 57, no. 6, p. 102124, 2020.
- [46] A. Alosaimy and E. Atwell, "Diacritization of a highly cited text: a classical arabic book as a case," in *2018 IEEE 2nd International Workshop on Arabic and Derived Script Analysis and Recognition (ASAR)*. IEEE, 2018, pp. 72–77.
- [47] A. Abdelli, F. Guerrouf, O. Tibermacine, and B. Abdelli, "Sentiment analysis of arabic algerian dialect using a supervised method," in *International Conf. on Intelligent Systems and Advanced Computing Sciences (ISACS)*. IEEE, 2019, pp. 1–6.
- [48] R. Moumen, R. Chiheb, R. Faizi, and A. El Afia, "Arabic diacritization with gated recurrent unit," in *Proceedings of the International Conf. on Learning and Optimization Algorithms: Theory and Applications*, 2018, pp. 1–4.
- [49] M. A. H. Madhfar and A. M. Qamar, "Effective deep learning models for automatic diacritization of arabic text," *IEEE Access*, vol. 9, pp. 273–288, 2020.
- [50] A. Fadel, I. Tuffaha, M. Al-Ayyoub *et al.*, "Arabic text diacritization using deep neural networks," in *2019 2nd International Conf. on Computer Applications & Information Security (ICCAIS)*. IEEE, 2019, pp. 1–7.
- [51] M. Maamouri, A. Bies, T. Buckwalter, and W. Mekki, "The Penn Arabic treebank: Building a large-scale annotated Arabic corpus," in *NEMLAR Conf. on Arabic Language Resources and Tools*, vol. 27. Cairo, 2004, pp. 466–467.
- [52] I. Zitouni, J. Sorensen, and R. Sarikaya, "Maximum entropy based restoration of Arabic diacritics," in *Proceedings of the 21st International Conf. on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, 2006, pp. 577–584.
- [53] M. M. Alamri and W. J. Teahan, "Automatic correction of Arabic dyslexic text," *Computers*, vol. 8, no. 1, p. 19, 2019.
- [54] Y. J. Choe, J. Ham, K. Park, and Y. Yoon, "A neural grammatical error correction system built on better pre-training and sequential transfer learning," *arXiv preprint arXiv:1907.01256*, 2019.

Prediction of Presence of Brain Tumor Utilizing Some State-of-the-Art Machine Learning Approaches

Mitrabinda Khuntia, Prabhat Kumar Sahu, Swagatika Devi

Dept. of Computer Science and Engineering, ITER, S'O'A University, Bhubaneswar, India

Abstract—A brain tumor is a kind of abnormal development caused by unregularized cell reproduction and it is increasing day-by-day. The Magnetic Resonance Imaging (MRI) tools are the most often used diagnostic tool for brain tumor detection. However, ample amount of information contained in MRI makes the detection and analysis process tedious and time consuming. The ability to accurately identify the exact size and proper location of a brain tumor is a tough task for radiologists. Medical image processing is an interdisciplinary discipline in which image processing is a tough research. Image segmentation is the prime requirement in image processing as it separates dubious regions from biomedical images thereby enhancing the treatment reliability. In this regard, our article reviews eight existing binary classifiers to compare their results for designing an automated Computer Aided Diagnosis (CAD) system. The proposed classification models can analyze T1-weighted brain MRI images to reach at a conclusion. The classification accuracy advocates the quality of our work.

Keywords—Brain tumor classification; MRI; SVM; decision tree; random forest; CAD

I. INTRODUCTION

The proper diagnosis of some crucial information is a key challenge in the field of bioinformatics or medical research. Many diagnostic and research institutions include a wealth of medical diagnosis data. It is barely essential to categorize them in order to automate and speed up illness diagnosis.

A continuous progression in cancer research has been carried out during the previous decades [1]. Scientists used a number of approaches, including very early-stage screening, to identify the disease before symptoms appeared. They have also developed noble methods for detecting the disease therapy results early on [16]. As a result of the advent of new medical technology, large amounts of cancer data have been gathered and made available to the clinical research community. Hence, medical researchers are exclusively employing popular machine learning techniques which can discover patterns and connections from massive datasets and anticipate future cancer outcomes with high accuracy.

The automated segmentation and categorization of medical images is crucial in brain tumor diagnosis, prognosis of tumor development, and therapy. Early diagnosis of a brain tumor predicts a faster response in therapy, which improves patient survival rates. Manual procedures used in normal clinical work to find and categorize brain tumors in large medical image collections incur considerable effort and time costs. It is desirable and beneficial to have a procedure for automatic detection, localization, and classification. Any disruption may intimate disease and injury. The identification of brain tumors is crucial in biological applications. Several procedures, like

as preprocessing, feature extraction, and classification, are required during the classification process.

Various medical imaging modalities are utilized to give tumor related information that is required for detection [2]. Prime methods include computed tomography (CT), single-photon emission computed tomography (SPECT), positron emission tomography (PET), magnetic resonance spectroscopy (MRS), and magnetic resonance imaging (MRI). MRI collection produces numerous 2D picture slices with strong tissue contrast while taking benefit of no ionising radiation [2]. T2 images are more suitable for identifying the borders of edoema areas. Brain tumors are identified and classified using MRI image processing.

II. LITERATURE SURVEY

Iftekharruddin et al.[5] used fractal wavelet characteristics as input to a Self Organizing Map (SOM) classifier and attained an average accuracy of 90%. Based on histogram study of temporal Magnetic Resonance Image (MRI) data, Manikis et al [14] developed a unique paradigm for monitoring tumor alterations. The proposed method detects tumor distribution and quantitatively predicts its development or decrease, possibly benefiting clinicians in objectively analysing tiny changes throughout therapy. Roy et al. [15] proposed an investigation towards automated brain tumor identification and classification using brain MRI. Brain tumor segmentation was a critical method for collecting information from complex MRI images of the brain. Sindhushree K.S et al [6] created and tested a strategy for segmenting brain tumors using two-dimensional MRI data. Discovered tumors are also shown in three dimensions. To identify malignancy, high pass filtering, histogram equalization, thresholding, morphological methods, along with segmentation employing linked component labeling were used. The recovered 2D tumor pictures were rebuilt into 3D volumetric data, and the tumor volume was determined.

Havaei et al.[6] designed a semi-automatic method using kNN classifier. They used the well-known BRATS 2013 dataset and done both whole and core tests, with Dice similarities of 0.85 for the total tumor region and 0.75 for the core tumor area. Sachdeva et al.[3] semi-automatically constructed the tumor contour and then calculated 71 features using the intensity profile, co-occurrence matrix, and Gabor functions. The classifiers Support Vector Machine (SVM) and Artificial Neural Network (ANN) were compared. Similarly, Kaur[4] presented autonomous brain tumor classification approach with ten features and a Back Propagation Neural Network as the classifier, which had a 95.3% accuracy.

Mohsen et al. [9] suggested a deep learning based classifier paired with discrete wavelet transform (DWT) and principal

components analysis to categorize a dataset including three distinct brain tumors (PCA). Four other deep learning-related research with equivalent goal employ the same dataset as we used in this study, which is crucial for comparing and evaluating the proposed model's performance outcomes. Pashaei et al. [11] suggested two approaches for classification: the first employed a CNN model for classification, while the second used CNN characteristics as inputs to a KELM methodology. The KELM algorithm is a learning algorithm composed of hidden node layers. A two-layer CNN design was introduced by Abiwinanda et al. [10]. A CNN with 16 convolution layers was proposed by Sultan et al. [12]. In another work, Anaraki et al. [13] introduced a hybrid approach for network design enhancement that combines the usage of CNNs with genetic algorithm (GA) criteria.

III. BRAIN TUMOR DIAGNOSIS

Brain is a centralized processing unit in humans that senses, controls, and runs all of our bodily functions. Neurons and Galilean cells are the two types of cells that make up the brain. Brain tumor refers to an unexpected proliferation of brain cells in the brain. Brain tumors can be either malignant or non-cancerous. The examination of tumors in the identification of malignant characteristics is a tough work owing to the variable nature of the tumor and its similarity to other regions of the brain. Early discovery of this impact has a higher possibility of recovery than late diagnosis. However, in today's world, the vast majority of tumors are identified at a late stage. As a result, early stage detection is a critical necessity.

IV. PROPOSED FRAMEWORK

A. Dataset Used

Benign cases are classed as positive in our study, whereas malignant ones are classified as negative as shown in Fig. 1 and fig. 2. Linear correlations are straight-line correlations between two variables with values ranging from -1 to + 1, where -1 represents the ideal negative relationship and + 1 represents the ideal positive relationship. By identifying the relationship between nine aspects of benign and malignant classes, the Pearson correlation between positive and negative classes is presented.

B. Block Diagram of the System

Before training the model, we collected images, partitioned the dataset, and investigated augmentation alternatives. The model is fine-tuned, and the outcomes were enhanced. The confusion matrix, model loss, and model accuracy have all been proven to show the loss and accuracy change with epoch. The proposed block diagram displays the whole classifier system in the simplest way possible as presented in Fig. 3. Decision making is a key component of this scheme and serves an important role in the research.

C. Data Preprocessing

Data preprocessing is used to fill in blanks, locate and eliminate outliers, and resolve self-contradiction. The sample code number has been removed from the dataset since it has no effect on illnesses. Vectors are created by resizing images.

They are then scaled to fit the training method [17]. The following step is to transform each image in the collection to an array. The image is used as a preprocessed input by MobileNetV2. The final level is coding. The tagged dataset is converted into a numerical label, which can then be understood and evaluated. Furthermore, random selection is used in the dataset to guarantee that the data is adequately disseminated.

D. Training and Testing

The training phase extracts properties from the dataset, while the testing phase assesses how well the appropriate model predicts. The dataset is divided into two sections. This is the time for training and testing. In K fold cross-validation, a single fold is utilized for testing and k-1 folds are used for training in a cyclical method. To avoid over fitting, cross-validation is performed. In this paper, we partition data using a five-fold cross-validation strategy, with four fold used for training and one-fold used for testing in each iteration.

E. Performance Measurements

Following the labeling of all pixels I_j in the input slice, P_{ij} as illustrated in equation (1) and (2). From vector f_l , $l = 1, 2, 3$, the classification function predicts the label l_p , that pinpoints the kind of tumor in a slice. The classification function determines the link between predicted label sizes, l , $P_{ij} == 1$, and the overall prediction, $P_{ij} > 0$. The projected label, l_p , will be the label with the largest capacity connection i.e. greater than the confidence threshold's minimum size relation, $\zeta_c \in [0, 1]$.

$$P_{ij} = \begin{cases} = 0, & \text{if (i,j) is healthy position} \\ = 1, & \text{if (i,j) is meningioma tumor} \\ = 2, & \text{if (i,j) is glioma tumor} \\ = 3, & \text{if (i,j) is pituitary tumor} \end{cases} \quad (1)$$

$$f_l = \begin{cases} \frac{P_{ij} == 1}{P_{ij} > 0} > \zeta_c \\ 0 \end{cases} \quad (2)$$

F. Image Classification Performance Metrics

Several metrics, including accuracy, precision, recall, F1-score, and AUC, were employed to evaluate performance of our scheme.

V. MACHINE LEARNING TECHNIQUES

When using traditional Machine Learning approaches, a preprocessing stage aimed at feature extraction is included in the segmentation pipeline [8]. The recovered attributes are then passed on to the classification or segmentation stage [7]. The ML inquiry would inquire whether or not the tumor is likely to be malignant (1=Yes, 0=No). Some important techniques to improve the performance of ML approaches are discussed below:

- 1) dimensionality reduction
- 2) feature selection
- 3) feature extraction

	0	1	2	3	4	5	6	7	8	9	...	150519	150520	150521	150522	150523
0	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	...	0.000000	0.000000	0.000000	0.000000	0.000000
1	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	...	0.000000	0.000000	0.000000	0.000000	0.000000
2	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	...	0.000000	0.000000	0.000000	0.000000	0.000000
3	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	...	0.000000	0.000000	0.000000	0.000000	0.000000
4	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	...	0.000000	0.000000	0.000000	0.000000	0.000000
...
248	0.140870	0.140870	0.140870	0.070387	0.070387	0.070387	0.062868	0.062868	0.062868	0.085950	...	0.078256	0.078256	0.078256	0.078256	0.078256
249	0.003922	0.003922	0.003922	0.003922	0.003922	0.003922	0.003922	0.003922	0.003922	0.003922	...	0.003922	0.003922	0.003922	0.003922	0.003922
250	0.143616	0.143616	0.143616	0.345293	0.345293	0.345293	0.301732	0.301732	0.301732	0.265334	...	0.272768	0.272768	0.272768	0.243786	0.243786
251	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	...	0.105882	0.105882	0.105882	0.108438	0.108438
252	0.119553	0.119553	0.119553	0.119553	0.119553	0.119553	0.119553	0.119553	0.119553	0.119553	...	0.080296	0.080296	0.080296	0.078648	0.078648

253 rows × 150529 columns

Fig. 1. Data Set of Brain Tumor

150524	150525	150526	150527	Target
0.000000	0.000000	0.000000	0.000000	0
0.000000	0.000000	0.000000	0.000000	0
0.000000	0.000000	0.000000	0.000000	0
0.000000	0.000000	0.000000	0.000000	0
0.000000	0.000000	0.000000	0.000000	0
...
0.078256	0.133088	0.133088	0.133088	1
0.003922	0.003922	0.003922	0.003922	1
0.243786	0.237173	0.237173	0.237173	1
0.108438	0.113270	0.113270	0.113270	1
0.078648	0.082783	0.082783	0.082783	1

Fig. 2. Segment of Fig. 1

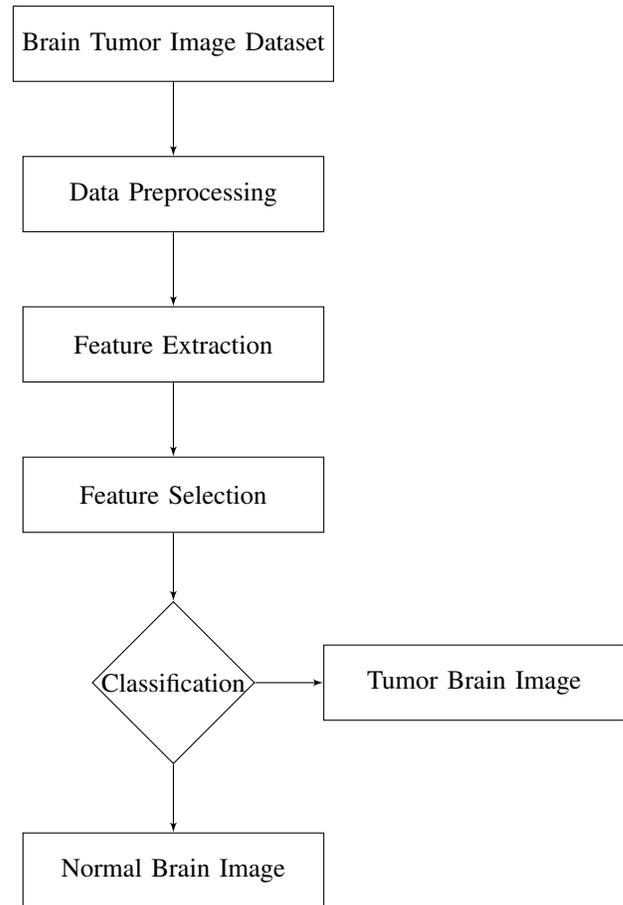


Fig. 3. Block Diagram

A. Support Vector Machine Algorithm

SVM or Support Vector Machine, may be employed for regression along with classification applications. SVMs offer much greater search accuracy than typical query refinement techniques after only 3 to 4 rounds of relevance feedback, according to simulated data. The SVM algorithm is frequently used in clinical and other disciplines. We will be utilizing the brain tumor dataset to develop our SVM method.

The F1-score for healthy and brain tumor categorization is 86% and 92%, respectively in Fig. 4. We can see from the

output of Fig. 4 that there were some inaccurate predictions; thus, if we want to determine the number of correct and incorrect predictions, we must utilize the confusion matrix. The confusion matrix is shown in the output graphic in Fig. 5, with 4+1=5 wrong guesses and 16+30=46 right predictions.

	precision	recall	f1-score	support
0	0.94	0.80	0.86	20
1	0.88	0.97	0.92	31
accuracy			0.90	51
macro avg	0.91	0.88	0.89	51
weighted avg	0.91	0.90	0.90	51

Fig. 4. Classification Metrics of SVM

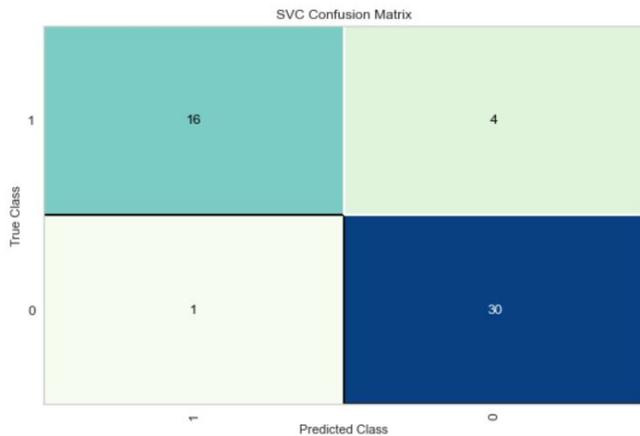


Fig. 5. Confusion Matrix for SVM Classifier

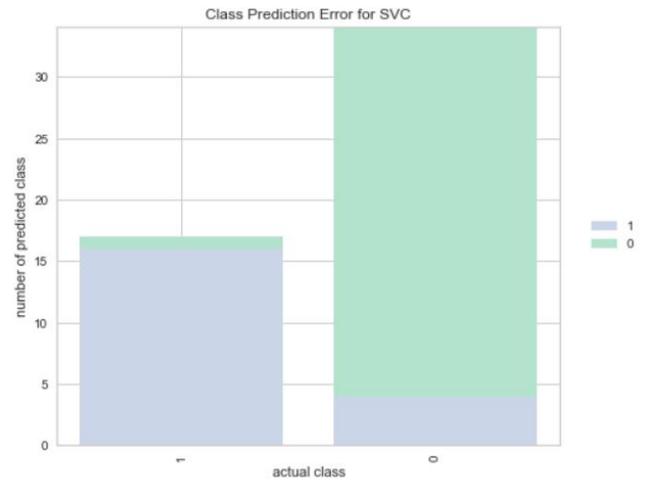


Fig. 7. Visualizer for SVM Classifier

	precision	recall	f1-score	support
0	0.76	0.65	0.70	20
1	0.79	0.87	0.83	31
accuracy			0.78	51
macro avg	0.78	0.76	0.77	51
weighted avg	0.78	0.78	0.78	51

Fig. 8. Classification Metrics for DTREE Classifier

The graphical depiction of the results in terms of ROC and micro-average ROC curve is shown in Fig. 6. We will draw a graph for the SVM classifier in Fig. 7 to illustrate the training set outcome. The classifier will determine whether a brain tumor is malignant or benign.

B. Decision Tree

Decision Tree (DTree) is a Supervised learning approach and tree structured that can be utilized to solve classification problems. Here the internal nodes represents attribute, branch represent rules and leaf node specifies conclusion. It is a

graphical depiction of solutions to a problem depending on specific criteria.

The F1-score for healthy and brain tumour categorization is 70% and 83%, respectively in Fig. 8. We can see from the output of Fig. 8 that there were some inaccurate predictions; thus, if we want to determine the number of correct and incorrect predictions, we must utilize the confusion matrix. The confusion matrix is shown in the output Fig. 9, with 4+7=11 inaccurate guesses and 13+27=40 right predictions.

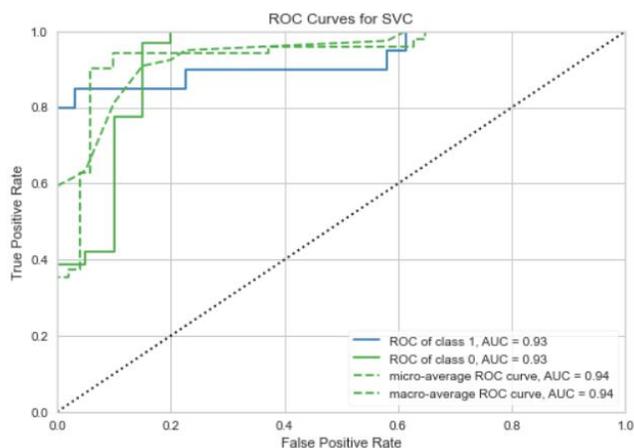


Fig. 6. ROC for SVM Classifier

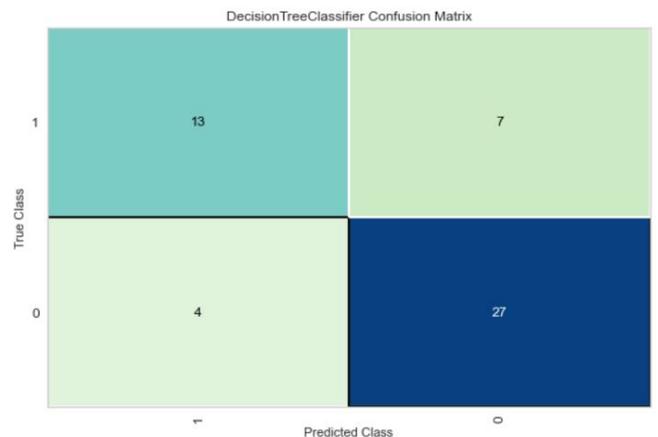


Fig. 9. Confusion Matrix for DTREE Classifier

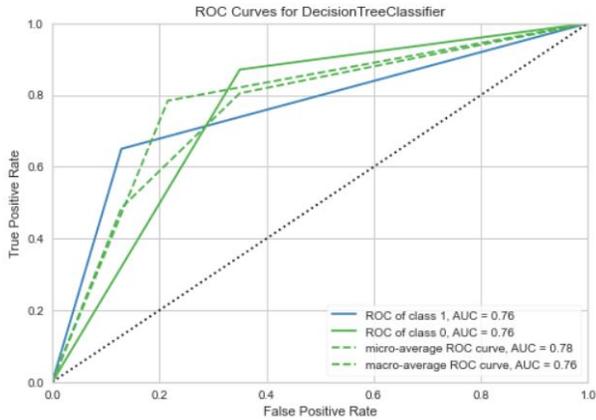


Fig. 10. ROC for DTREE Classifier

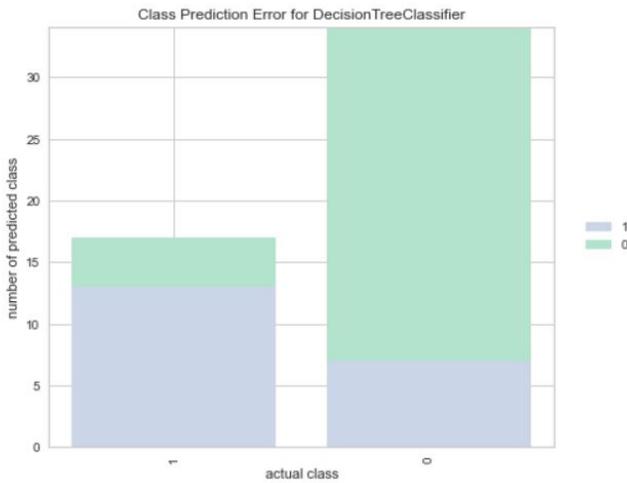


Fig. 11. Visualizer for DTREE Classifier

The graphical depiction of the results in terms of ROC and micro-average ROC curve is shown in Fig. 10. The above output is completely different from the rest classification models. We will plot a graph for the decision tree classifier in Fig. 11 to view the training set outcome. The classifier will determine whether a brain tumor is malignant or benign.

C. Gaussian Naive Bayes

Gaussian Naive Bayes is the name given to the generalization of naive Bayes. The normal distribution (Gaussian distribution) is simpler to use as it can estimate mean and standard deviation very quickly from the training data. A Gaussian distribution is assumed if the input variables are real-valued. This may need the removal of outliers.

The F1-score for normal and brain tumor categorization is 81% and 87%, respectively which is represented in Fig. 12. The confusion matrix is shown in the output Fig. 13, with 3+5=8 inaccurate guesses and 17+26=43 right predictions.

The graphical depiction of the results in terms of ROC and micro-average ROC curve is shown in Fig. 14. We will draw

	precision	recall	f1-score	support
0	0.77	0.85	0.81	20
1	0.90	0.84	0.87	31
accuracy			0.84	51
macro avg	0.83	0.84	0.84	51
weighted avg	0.85	0.84	0.84	51

Fig. 12. Classification Metrics for Gaussian NB Classifier

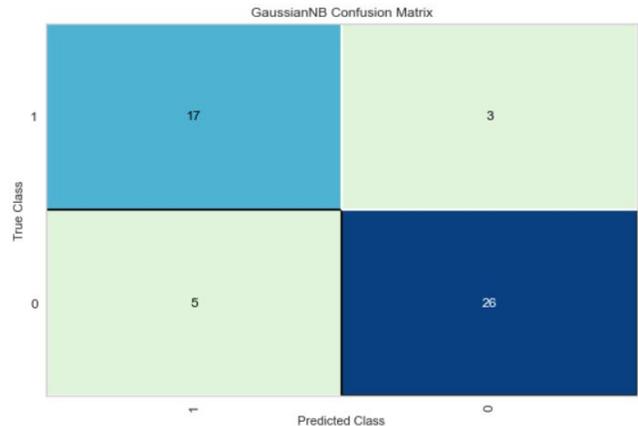


Fig. 13. Confusion Matrix for Gaussian NB Classifier

a graph in Fig. 15 for the Gaussian NB classifier to show the training set outcome. The classifier will determine whether a brain tumor is malignant or benign.

D. Random Forest

Random forest chooses observations at random, creates a decision tree, and uses the average result. Random Forest classifiers can handle both categorized as well as continuous variables in an effective manner. It outperforms other algorithms in categorization tasks. It is capable of handling binary,

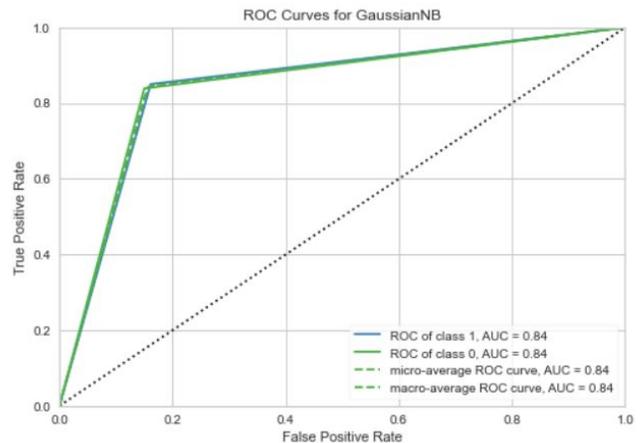


Fig. 14. ROC for Gaussian NB Classifier

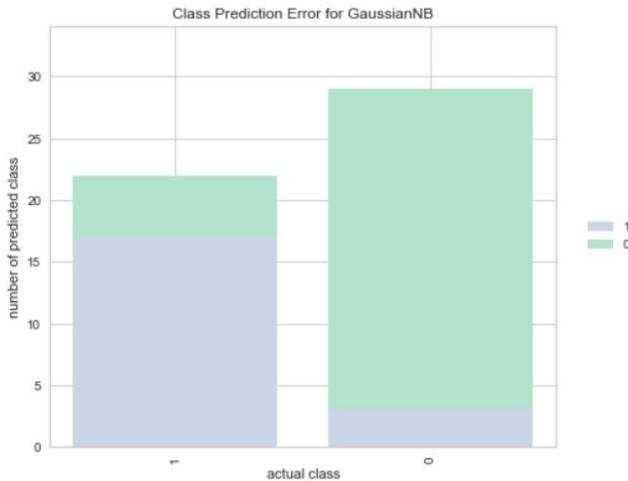


Fig. 15. Visualizer for Gaussian NB Classifier

	precision	recall	f1-score	support
0	1.00	0.85	0.92	20
1	0.91	1.00	0.95	31
accuracy			0.94	51
macro avg	0.96	0.93	0.94	51
weighted avg	0.95	0.94	0.94	51

Fig. 16. Classification Metrics for Random Forest Classifier

continuous, and categorical data.

The F1-score for healthy and brain tumor categorization is 92% and 95%, respectively in Fig. 16. The confusion matrix is constructed in the output image Fig. 17 to identify the accurate and wrong guesses, which contains 3+0=3 erroneous predictions and 17+31=48 correct predictions.

The graphical depiction of the results in terms of ROC and micro-average ROC curve is shown in Fig. 18. The above output is completely different from the rest classification

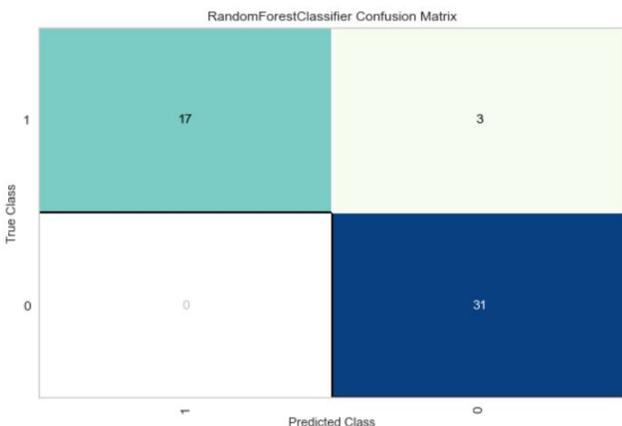


Fig. 17. Confusion Matrix for Random Forest Classifier

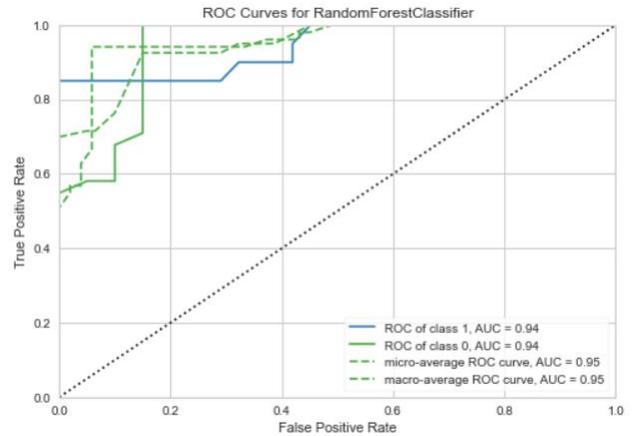


Fig. 18. ROC for Random Forest Classifier

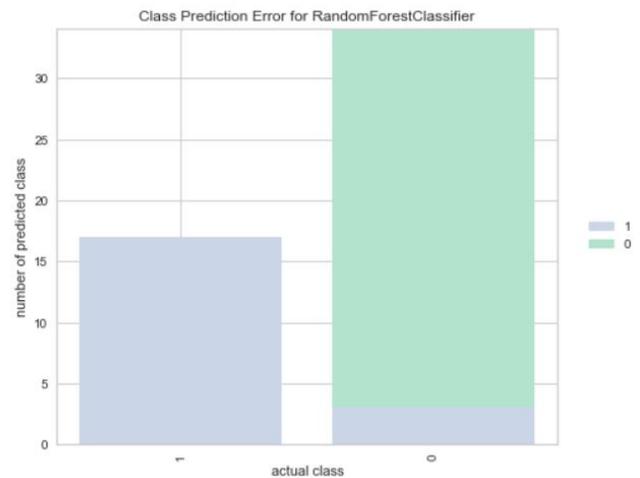


Fig. 19. Visualizer for Random Forest Classifier

models. We will create a graph for the Random Forest classifier to view the training set results in Fig. 19. The classifier will determine whether a brain tumor is malignant or benign.

E. Multinomial Naive Bayes Classifier

Multinomial Naive Bayes can be treated as a probabilistic process and is extensively utilized for categorical training set. It helps in obtaining highest likelihood. Normally, the multinomial distribution requires integer feature counts.

The F1-score for normal and brain tumor categorization is

	0	0.78	0.70	0.74	20
1	0.82	0.87	0.84		31
accuracy				0.80	51
macro avg	0.80	0.79	0.79		51
weighted avg	0.80	0.80	0.80		51

Fig. 20. Classification Metrics for Multinomial NB Classifier

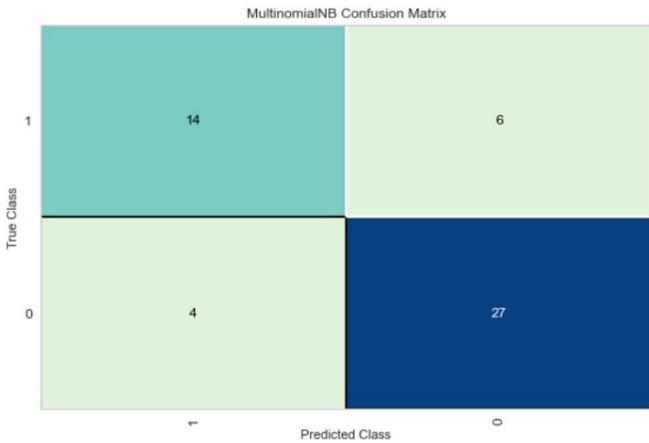


Fig. 21. Confusion Matrix for Multinomial NB Classifier

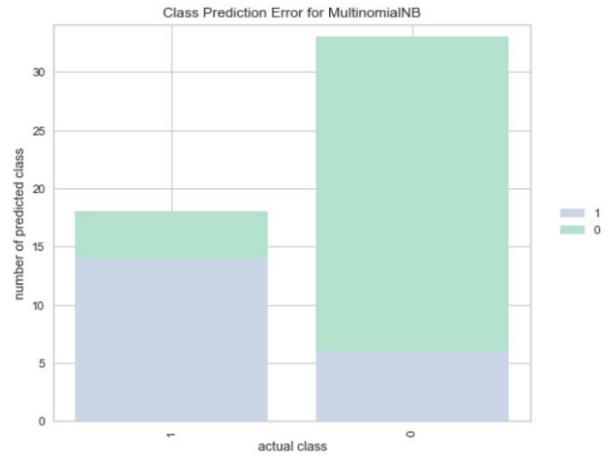


Fig. 23. Visualizer for Multinomial NB Classifier

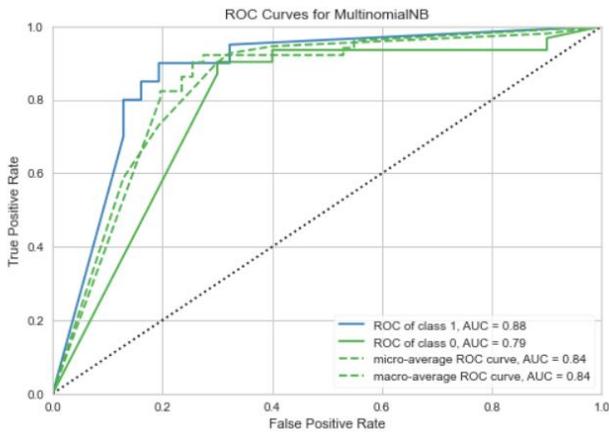


Fig. 22. ROC for Multinomial NB Classifier

	precision	recall	f1-score	support
0	1.00	0.85	0.92	20
1	0.91	1.00	0.95	31
accuracy			0.94	51
macro avg	0.96	0.93	0.94	51
weighted avg	0.95	0.94	0.94	51

Fig. 24. Classification Metrics for Extreme Gradient Boost Classifier

74% and 84%, respectively in Fig. 20. In Fig. 21, a confusion matrix is used to determine the correct and incorrect guesses, with $6+4=10$ incorrect forecasts and $14+27=41$ accurate predictions.

The graphical depiction of the results in terms of ROC and micro-average ROC curve is shown in Fig. 22. The above output is completely different from the rest classification models. We will draw a graph of Fig. 23 for the Multinomial NB classifier to show the training set outcome. The classifier will determine whether a brain tumor is malignant or benign.

F. Extreme Gradient Boost (XGB) Classifier

This classifier is treated as a boosted classifier for tabular as well as structured training samples. At the same time, it has the characteristic to handle complex and huge databases. It is a technique for ensemble modeling.

The F1-score for healthy and brain tumor categorization is 92% and 95%, respectively in Fig. 24. In Fig. 25, a confusion matrix is used to determine the correct and incorrect guesses, with $3+0=3$ erroneous forecasts and $17+31=48$ accurate predictions.

The graphical depiction of the results in terms of ROC and micro-average ROC curve is shown in Fig. 26. To display the training set outcome, we will create a graph for the XGB classifier in Fig. 27. The classifier will determine whether the brain tumor is malignant or benign. XGBoost is more than 10 times quicker.

G. Stochastic Gradient Descent (SGD) Classifier

A gradient is the slope of a function. It assesses the degree to which one variable changes in response to changes in

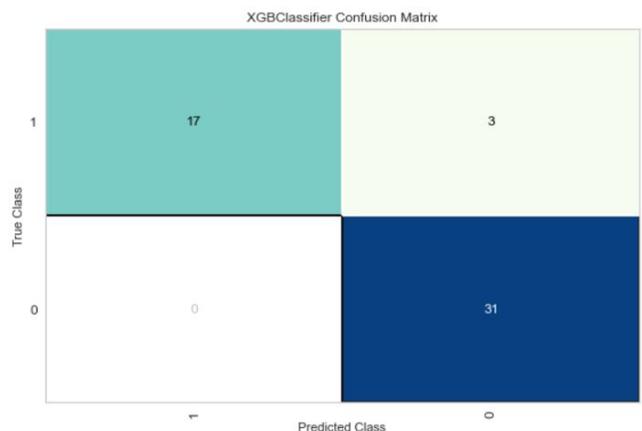


Fig. 25. Confusion Matrix for XGB Classifier

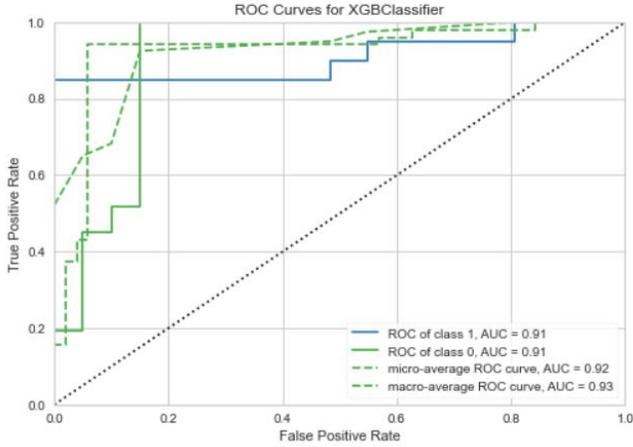


Fig. 26. ROC for XGB Classifier

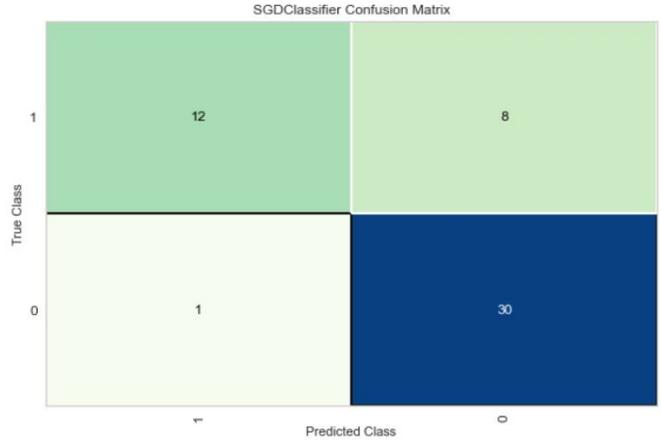


Fig. 29. Confusion Matrix for SGD Classifier

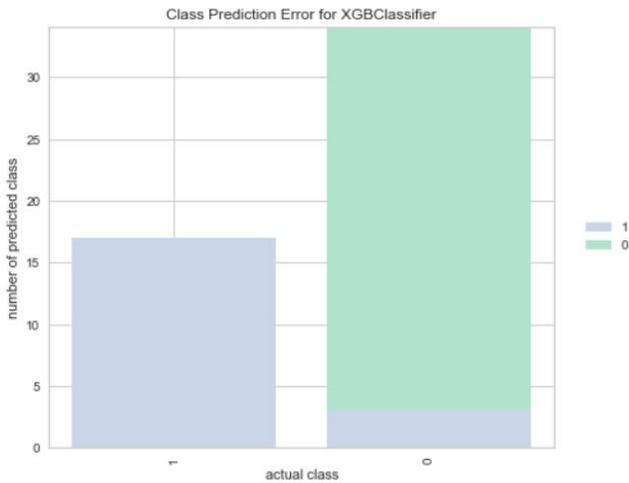


Fig. 27. Visualizer for XGB Classifier

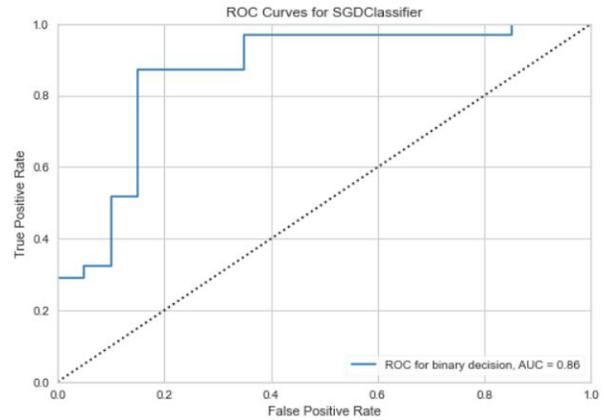


Fig. 30. ROC for SGD Classifier

another one. The steeper the slope, the higher the gradient value. It computes gradient using a single training sample. It is quicker and less computationally costly than batch gradient descent.

The F1-score for healthy and brain tumor categorization is 73% and 87%, respectively in Fig. 28. In the graph of Fig. 29, a confusion matrix is used to determine the correct and incorrect guesses, with $8+1=9$ incorrect forecasts and $12+30=42$ accurate predictions.

	precision	recall	f1-score	support
0	0.92	0.60	0.73	20
1	0.79	0.97	0.87	31
accuracy			0.82	51
macro avg	0.86	0.78	0.80	51
weighted avg	0.84	0.82	0.81	51

Fig. 28. Classification Metrics for SGD Classifier

The graphical depiction of the results in terms of ROC and micro-average ROC curve is shown in Fig. 30. The above output is completely different from the rest classification models. We will plot a graph for the SGD classifier in Fig. 31 to visualize the training set outcome. The classifier will determine whether a brain tumor is malignant or benign.

H. Bagging Classifier

Bagging lowers over fitting (variance) by averaging or voting; nevertheless, this increases bias, which is offset by the decrease in variance. Bagging builds n classification trees from the training data using bootstrap sampling and then combines their predictions to get a final meta-prediction. Bagging and decision trees can be combined and used to eliminate overfitting.

The F1-score for healthy and brain tumor categorization is 90% and 94%, respectively in Fig. 32. In the graphic of Fig. 33, a confusion matrix is used to calculate the correct and incorrect guesses, with $2+2=4$ incorrect predictions and $18+29=47$ accurate predictions.

The graphical depiction of the results in terms of ROC and micro-average ROC curve is shown in Fig. 34. The above

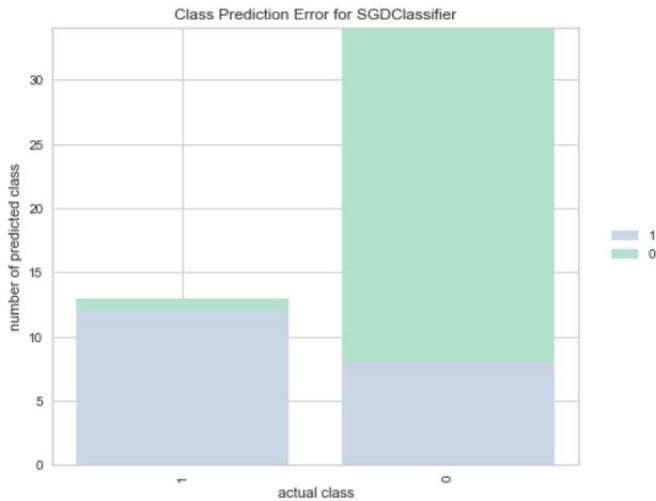


Fig. 31. Visualizer for SGD Classifier

	precision	recall	f1-score	support
0	0.90	0.90	0.90	20
1	0.94	0.94	0.94	31
accuracy			0.92	51
macro avg	0.92	0.92	0.92	51
weighted avg	0.92	0.92	0.92	51

Fig. 32. Classification Metrics for BAG Classifier

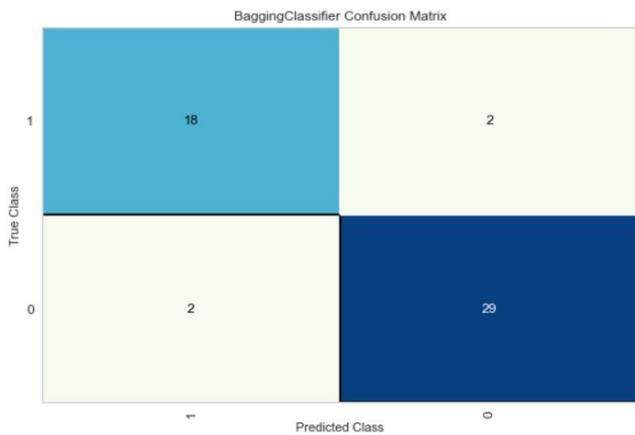


Fig. 33. Confusion Matrix for BAG Classifier

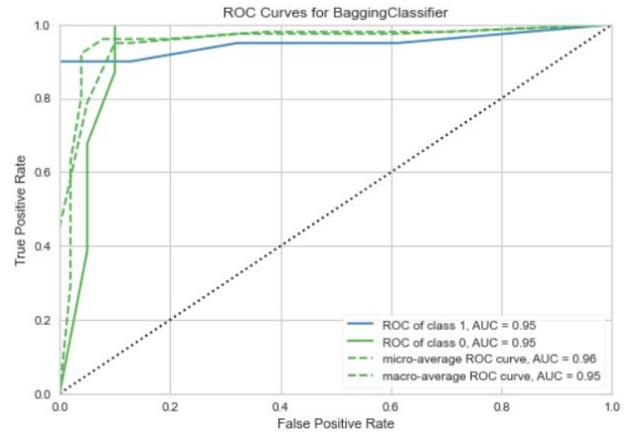


Fig. 34. ROC for BAG Classifier

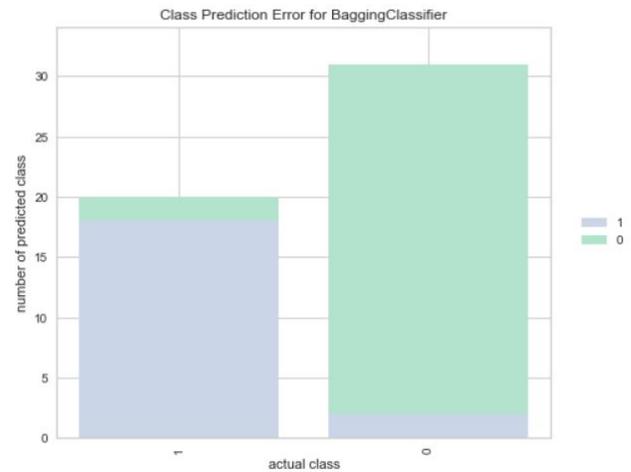


Fig. 35. Visualizer for BAG Classifier

output is completely different from the rest classification models. We will plot a graph for the BAG classifier in Fig. 35 to visualize the training set outcome. The classifier will determine whether a brain tumor is malignant or benign.

I. LGBM Classifiers

Light GBM can handle enormous quantities of data while consuming minimal memory. It emphasises result precision. LGBM also supports GPU learning, therefore scientists are

	precision	recall	f1-score	support
0	0.94	0.80	0.86	20
1	0.88	0.97	0.92	31
accuracy			0.90	51
macro avg	0.91	0.88	0.89	51
weighted avg	0.91	0.90	0.90	51

Fig. 36. Classification Metrics for LGBM Classifier

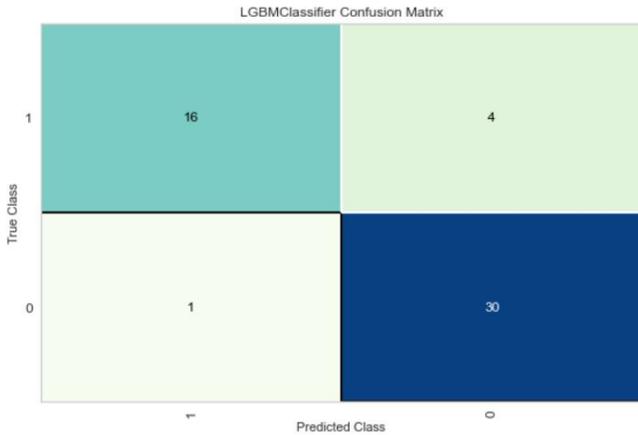


Fig. 37. Confusion Matrix for LGBM Classifier

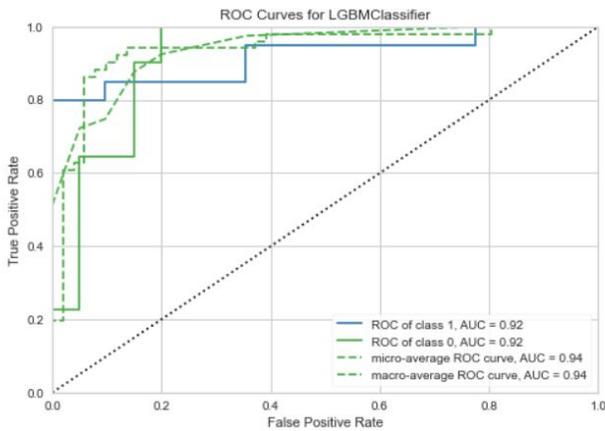


Fig. 38. ROC for LGBM Classifier

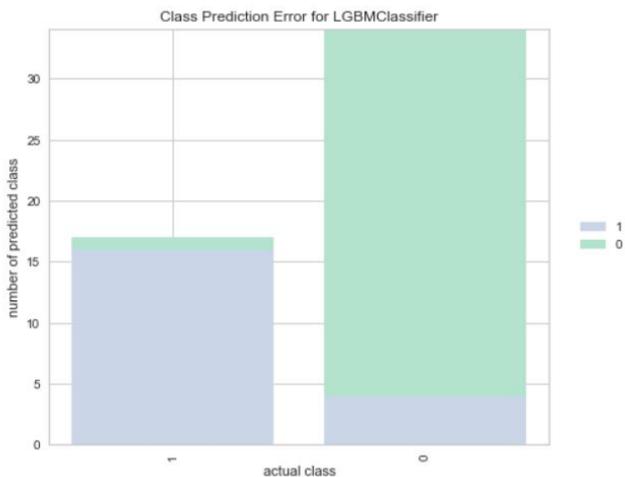


Fig. 39. Visualizer for LGBM Classifier

	model	best_score	best_params
0	svm	0.747805	{'C': 30, 'kernel': 'rbf'}
1	random_forest	0.772561	{'n_estimators': 10}
2	logi_reg	0.757805	{'C': 1}
3	gNB	0.757805	{'var_smoothing': 0.02310129700083159}
4	mNB	0.688293	{'alpha': 0.9, 'fit_prior': True}
5	bGC	0.816829	{'max_features': 30, 'n_estimators': 70}
6	dDC	0.757195	{'criterion': 'gini', 'max_depth': 5, 'min_sam...
7	Xgb	0.801829	{'colsample_bytree': 0.8, 'gamma': 0, 'learnin...
8	SGD	0.743293	{'loss': 'hinge', 'penalty': 'l2'}

Fig. 40. Comparison of different Techniques of Machine Learning for the Prediction of Brain Cancer in GridSearchCV

	model	best_score	best_params
0	svm	0.747805	{'kernel': 'rbf', 'C': 30}
1	random_forest	0.787073	{'n_estimators': 30}
2	logi_reg	0.757805	{'C': 1}
3	gNB	0.757805	{'var_smoothing': 0.02310129700083159}
4	mNB	0.688293	{'fit_prior': True, 'alpha': 0.9}
5	bGC	0.821829	{'n_estimators': 70, 'max_features': 10}
6	dDC	0.757195	{'min_samples_leaf': 5, 'max_depth': 5, 'crite...
7	Xgb	0.801829	{'subsample': 0.8, 'seed': 27, 'scale_pos_weig...
8	SGD	0.753049	{'penalty': 'elasticnet', 'loss': 'log'}

Fig. 41. Comparison of different Techniques of Machine Learning for the Prediction of Brain Cancer in RandomizedSearchCV

utilizing it to build research applications. LGBM should not be used to small datasets.

The F1-score for healthy and brain tumor classification is 86% and 92%, respectively in Fig. 36. A confusion matrix is used in the graph of Fig. 37 to calculate the correct and incorrect guesses, with 4+1=5 incorrect predictions and 16+30=46 accurate predictions.

The graphical depiction of the results in terms of ROC and micro-average ROC curve is shown in Fig. 38. The above output is completely different from the rest classification models. To illustrate the training set outcome, we shall draw a graph for the LGBM classifier in Fig. 39. The classifier will evaluate whether a brain tumor is benign or malignant.

VI. EXPERIMENTAL ANALYSIS AND RESULT

We compared all the techniques used for the prediction of brain cancer by different parameters in grid search CV (Fig. 40) and randomized search CV (Fig. 41), respectively. The proposed method was implemented in Python by using 5-fold cross validation techniques. Our experimental result proves that all the nine classifier are providing good results with respect to different parameters values. However, for both GridSearchCV and RandomizedSearchCV bagging classifier is giving best results.

VII. CONCLUSION

Our article employs data augmentation approach prior to classification to avoid overfitting. We surveyed some popular state-of-the-art machine learning approaches to reach at a conclusion. Our work is experimented on T1-weighted contrast-enhanced MRI images. However, this study reveals the importance of supervised learning approaches on devising CAD systems to reduce the burden of radiologists. A future exploration can be extended in collecting some larger brain MR images to generalize the classifier systems.

REFERENCES

- [1] Hanahan, D., and Weinberg, R. A. (2011). Hallmarks of cancer: the next generation. *cell*, 144(5), 646-674.
- [2] Işın, A., Direkkoğlu, C., and Şah, M. (2016). Review of MRI-based brain tumor image segmentation using deep learning methods. *Procedia Computer Science*, 102, 317-324.
- [3] Sachdeva, J., Kumar, V., Gupta, I., Khandelwal, N., & Ahuja, C. K. (2016). A package-SFERCB-"Segmentation, feature extraction, reduction and classification analysis by both SVM and ANN for brain tumors". *Applied soft computing*, 47, 151-167.
- [4] Kaur, N., & kaur Panag, N. (2016). A Review on Brain tumour segmentation.
- [5] Iftekharuddin, K. M., Zheng, J., Islam, M. A., & Ogg, R. J. (2009). Fractal-based brain tumor detection in multimodal MRI. *Applied Mathematics and Computation*, 207(1), 23-41.
- [6] Havaei, M., Jodoin, P. M., & Larochelle, H. (2014, August). Efficient interactive brain tumor segmentation as within-brain kNN classification. In *2014 22nd international conference on pattern recognition* (pp. 556-561). IEEE.
- [7] Bishop, C. M., & Nasrabadi, N. M. (2006). *Pattern recognition and machine learning* (Vol. 4, No. 4, p. 738). New York: springer.
- [8] Chaddad, A., & Tanougast, C. (2016). Quantitative evaluation of robust skull stripping and tumor detection applied to axial MR images. *Brain informatics*, 3(1), 53-61.
- [9] Mohsen, H., El-Dahshan, E. S. A., El-Horbaty, E. S. M., & Salem, A. B. M. (2018). Classification using deep learning neural networks for brain tumors. *Future Computing and Informatics Journal*, 3(1), 68-71.
- [10] Abiwinanda, N.; Hanif, M.; Hesaputra, S.T.; Handayani, A.; Mengko, T.R. Brain tumor classification using convolutional neural network. In *Proceedings of the World Congress on Medical Physics and Biomedical Engineering, Prague, Czech Republic, 3-8 June 2018*; Springer: Singapore, 2019; pp. 183-189.
- [11] Pashaei, A., Sajedi, H., & Jazayeri, N. (2018, October). Brain tumor classification via convolutional neural network and extreme learning machines. In *2018 8th International conference on computer and knowledge engineering (ICCKE)* (pp. 314-319). IEEE.
- [12] Sultan, H. H., Salem, N. M., & Al-Atabany, W. (2019). Multi-classification of brain tumor images using deep neural network. *IEEE Access*, 7, 69215-69225.
- [13] Anaraki, A. K., Ayati, M., & Kazemi, F. (2019). Magnetic resonance imaging-based brain tumor grades classification and grading via convolutional neural networks and genetic algorithms. *biocybernetics and biomedical engineering*, 39(1), 63-74.
- [14] Manikis, G. C., Emmanouilidou, D., Sakkalis, V., Graf, N., & Marias, K. (2011, November). A fully automated image analysis framework for quantitative assessment of temporal tumor changes. In *2011 E-Health and Bioengineering Conference (EHB)* (pp. 1-6). IEEE.
- [15] Roy, S., Nag, S., Maitra, I. K., & Bandyopadhyay, S. K. (2013). A review on automated brain tumor detection and segmentation from MRI of brain. *arXiv preprint arXiv:1312.6150*.
- [16] Wang, T., Manohar, N., Lei, Y., Dhabaan, A., Shu, H. K., Liu, T., ... & Yang, X. (2019). MRI-based treatment planning for brain stereotactic radiosurgery: dosimetric validation of a learning-based pseudo-CT generation method. *Medical Dosimetry*, 44(3), 199-204.
- [17] Kotte, S., Pullakura, R. K., & Injeti, S. K. (2018). Optimal multilevel thresholding selection for brain MRI image segmentation based on adaptive wind driven optimization. *Measurement*, 130, 340-361.

Mining Hidden Partitions of Voice Utterances using Fuzzy Clustering for Generalized Voice Spoofing Countermeasures

Sarah Mohammed Altuwayjiri¹, Ouiem Bchir², Mohamed Maher Ben Ismail³

Department of Computer Science,
College of Computer and Information Sciences,
King Saud University,
Riyadh 11543, Saudi Arabia^{1,2,3}
Department of Computer Science,
College of Computing and Informatics,
Saudi Electronic University,
Riyadh 11673, Saudi Arabia¹

Abstract—The high level of usability achieved by voice biometrics compared to other biometric authentication modalities has promoted the widespread use of automatic speaker verification (ASV) systems as authentication tools for several services in various domains. Despite their satisfactory performance, ASV systems are vulnerable to malicious voice spoofing attacks. Hence, voice spoofing countermeasures have emerged as essential solutions to stop such harmful attacks and protect ASV systems as well as users' confidentiality. Typically, these countermeasures classify utterances into genuine and spoofing categories. In this research, we propose two voice spoofing countermeasures that mainly aim to improve the generalization of supervised learning models. This goal is achieved through the adaptive handling of the high variance of both utterance classes, i.e., genuine and spoofing classes. The proposed spoofing countermeasure addresses the poor generalization problem by identifying the hidden structure of each utterance category prior to the classification task. Specifically, fuzzy clustering algorithms were deployed to mine the hidden partitions of utterance classes. The conducted experiments showed that the proposed approach outperforms the state-of-the-art approaches in the ASVspoof 2017 dataset, with a testing EER equal to 1.07%.

Keywords—Voice spoofing; spoofing countermeasure; classification; clustering

I. INTRODUCTION

At present, biometric authentication along with other identification features is widely deployed to manage, administrate and control systems' accessibility in order to secure the applications and stored data [43]. In particular, the widespread use of biometric recognition systems has prompted research efforts to consider various modalities such as retinal, facial and speech data. Speaker verification (SV) has been introduced as a biometric recognition paradigm that uses human voiceprints to identify individuals every time they access a given service or system. SV-based identification is typically meant to compare the speaker's voice with the voiceprints previously recorded, then grant access to the identified persons only. The advent of voice assistant and smart home devices boosted the interest in automatic speaker verification (ASV) systems as promising alternatives to ensure the security of various smart

home applications, smart devices, online payment processes and phone banking [30]. However, these ASV systems have proven to be vulnerable to voice spoofing attacks [41]. In fact, such attacks have been defined as presentation attacks (PA) according to the International Organization for Standardization (ISO) and the International Electrotechnical Commission (IEC) [15]. In fact, spoofing attacks occur when a fraudster falsifies another identity to access some personal or secured resources [26]. For instance, they can be achieved through replay attacks which consist of collecting voice samples of a particular person, manipulating them to produce a spoofing voice, and replaying the resulting spoofing voice to mislead an automatic speaker verification (ASV) system. This kind of voice manipulation can be performed using data voice conversion or speech synthesis algorithms [41, 48]. Obviously, in order to prevent these spoofing attacks, audio anti-spoofing countermeasures are required. A voice spoofing countermeasure is a classification system that can automatically categorize voice records into two predefined categories: genuine and spoofing. Typically, it comprises two main components: (i) an audio feature extractor, and (ii) a supervised learning model. In this context, various audio features have been investigated for designing highly discriminative descriptors. Namely, the constant Q cepstral coefficients (CQCC) [45] and the linear frequency cepstral coefficient (LFCC) [42] were proposed to better discriminate between spoofing and genuine voice records. Similarly, diverse classifiers, such as the Gaussian mixture models (GMMs) [39] and deep neural networks (DNNs) [36] have been extensively used to build models that can accurately map unseen voice records into the two predefined classes. Although different deep learning architectures, such as the residual neural networks (ResNets) [22] and the recurrent neural networks (RNNs) [9] have been adapted and used for anti-spoofing, GMM-based solutions overtake the state-of-the-art anti-spoofing recognition systems [41]. Nevertheless, one of the major unsolved issues affecting the reliability and accuracy of anti-spoofing recognition systems is the poor generalization of the learned models [41]. In other words, the learned model failed to predict unseen data instances. Generalization characterizes the model's ability to predict unseen data instances.

This limitation can be attributed to the high variance in both the genuine and the spoofed utterances. Specifically, genuine speech instances exhibit high interspeaker variance, owing to the discrepancies between speaker voices, as well as an intraspeaker variance because of the inconsistency in the human voice that can be affected by aspects such as the emotional state [1, 29]. These two types of variations also apply to spoofed speeches. Moreover, spoofing utterances witnessed other types of variations caused by the recording devices used to collect the original voice records and the algorithms used to manipulate them [1, 2, 9, 15, 22, 26, 29, 36, 39, 41, 42, 45, 48]. Hence, better handling of the high variance of the genuine and spoofed classes would improve the generalization of the spoofing countermeasures and, consequently, enhance the detection performance. In this paper, we propose to improve the recognition of voice spoofing utterances by tackling the problem of intraclass variation for two categories: genuine and spoofed. More specifically, we propose learning the underlying structure of each category (genuine/spoof) by clustering them into homogeneous sub-categories. The rest of the paper is organized as follows: In Section II, existing voice-spoofing countermeasures are surveyed. In Section III, background knowledge on clustering techniques is provided. The proposed approach is presented in Section IV. The experiments conducted are described in Section V along with the reported results and their analysis.

II. LITERATURE REVIEW

Recently, several spoofing countermeasures have been proposed. The development of these systems was boosted by contests in 2015 and 2017 [30, 43], which provided challenging data for anti-spoofing systems. More specifically, the training set contains five types of spoofing attack algorithms, referred to as known attacks, whereas the evaluation set, used for testing, contains the known attacks and five more types of attacks called unknown attacks. The proposed system amounts to classification systems for genuine and spoofed utterances. They aimed to discriminate spoofing utterances from genuine utterances. One of the main aspects that has been exploited is the presence of noise in the spoofing records [23]. During the playback and re-recording phases used by the replay attack, different types of noise are generated. These types of noise are mainly from the recording environment and recording device. They can potentially allow for differentiation between spoofed and genuine signals. In this context, both conventional and deep learning approaches have been reported.

A. Conventional Countermeasures

Typically, conventional spoofing/genuine recording classification comprises a feature extraction component followed by a classification component. The system proposed in [23] extracts the cepstral coefficient (CQCCs) [7] feature. A GMM [5] classifier was employed for the classification task. This system has been considered a baseline approach for recently proposed research for evaluating anti-spoofing systems [30, 43]. Alternatively, the system reported in [39] combines cochlear filter cepstral coefficients (CFCC) [33] and the instantaneous frequency (IF) [40]. The combined feature aims to capture the speech synthesis and voice conversion, thus characterizing spoofing utterances. It was then fed to the GMM classifier.

TABLE I. SUMMARY OF CONVENTIONAL APPROACHES FOR SPOOFING / GENUINE CLASSIFICATION

Reference	Feature	Classifier	Dataset	Training EERor rate (%)	Testing EERor rate (%)
[45]	CQCCs	GMM	ASVspoof 2015 [49]	0.048	0.462
[27]	CQCCs	GMM	ASVspoof 2017 [27]	10.35	24.77
[39]	MFCC CFC-CIF	GMM	ASVspoof 2015 [49]	0.408	2.013
[42]	LFCC	GMM, SVM	ASVspoof 2015 [49]	0.11	1.67
[37]	MFCC, MFPC, CosPhasePCs	SVM with i-vectors	ASVspoof 2015 [49]	0.008	3.922

Similarly, the work in [42] focused on segregating spoofing records generated by voice conversion or speech synthesis algorithms. For this purpose, the authors in [42] conducted an empirical comparison of 19 different features to determine the most appropriate one for classifying spoofing versus genuine records. These features are then conveyed to both GMM [5] and SVM [6] classifiers. The experimental results reported in [42] showed that the system comprising the LFCC [53] feature extraction component and GMM classifier component outperformed all other considered systems. On the other hand, the work in [37] used SVM as a classifier [6] for different extracted features. In addition, the i-vector was used for each feature and then integrated into a one-centered i-vector with a normalized length. The experimental results in [37] showed that the MFCC [11], Mel-frequency principal coefficients (MFPC) [14], and CosPhase principal coefficients (CosPhasePC) [47] fed into the SVM classifier had better classification performance. Table I presents a brief summary of conventional approaches for spoofing/genuine classification. All of these systems have experimented on the ASVspoof 2015 dataset [49].

B. Deep Learning based Countermeasures

The successful achievement of deep neural networks (DNN) in classification tasks has motivated the application of such approaches for anti-spoofing. Recently, deep learning approaches have been proposed for voice spoofing classification. In particular, the residual network model (ResNet) found recent success in the works of [8, 30, 36]. Moreover, as voice spoofing data can be considered as a sequence classification task, recurrent neural networks (RNNs) [35] were also investigated in the works [9, 19, 31, 51, 52]. Furthermore, while raw audio data were considered as inputs to the DNN model in some studies [9, 19, 30], engineered features were considered in others [8, 31, 36, 52].

1) *Residual Network based Approaches*: The proposed system in [30] employs a dilated residual network (DRN) deep-learning architecture [21]. The latter is based on the ResNet model, and an attention-filtering mechanism. More precisely, the DRN uses convolution layers instead of fully connected layers, and alters the residual units by adding a dilation factor. The attention component aims to select important parts while ignoring unrelated ones, such as the background noise segments [50]. Similarly, the system proposed in [8] employed the ResNet [21] deep-learning model. However, the proposed approach applies a deep-learning architecture in conjunction with two low-level cepstral features. In fact, the input conveyed

TABLE II. SUMMARY OF DEEP LEARNING APPROACHES FOR SPOOFING / GENUINE CLASSIFICATION

Reference	Feature	Classifier	Dataset	Training EERor rate (%)	Testing EERor rate (%)
[30]	Signal Logspec via FFT	ResNet	ASVspoof 2017[27]	6.09	8.54
[8]	CQCC and MFCC	GMM, ResNet	ASVspoof 2017[27]	2.58	13.30
[36]	Fusion of HFCC and CQCC	DNN, SVM	ASVspoof 2017[27]	7.6	11.5
[9]	MFCC, Fbank	LSTM and GRU	ASVspoof 2017[27]	6.32	9.81
[31]	CQT and FFT	RNN, LCNN, SVM, CNN + RNN	ASVspoof 2017[27]	3.95	6.73
[19]	Fbank	CNN + RNN (GRU)	ASVspoof 2015 [49]	0.03	1.97
[52]	Spectrogram features	CNN + RNN	ASVspoof 2015 [49]	0.40	3.33

to the network is not raw audio data but features extracted by MFCCs [11] and CQCCs [45]. Moreover, a GMM classifier is used at the back end of the network. The work in [36] uses a similar model, but it exploits high-frequency cepstral coefficients (HFCCs) instead of MFCCs [11] at the input of the network.

2) *Recurrent Neural Network based Approaches*: The authors of [9] used recurrent neural networks (RNN) [35] for spoofing/genuine record classification. More specifically, the proposed system employs long short-term memory (LSTM) [24] and a gated recurrent unit (GRU) [10]. LSTM has also been used in [44], where the proposed architecture consists of multiple dense layers followed by one or more LSTM layers. Similarly, the works in [19, 52] exploited the RNN [35] deep-learning model. However, an RNN is used with a convolutional neural network (CNN) [20]. More precisely, CNN is used as a feature extractor and RNN is used for processing long dependencies. The work in [19] crops the input records and trains the two models separately and uses a linear discriminant analysis (LDA) [25] as a back-end classifier. However, the work in [52] uses an end-to-end model. It uses a context window for input, and trains both models simultaneously by conjointly optimizing them through backpropagation. The combination of CNNs and RNNs was also exploited in [31]. Specifically, it fuses three approaches: the i-vector [12] approach, light convolutional neural network (LCNN) [46] approach, and CNN+RNN approach. Three inputs were considered separately in the first convolution layer yielding three variants of the LCNN-based model. More precisely, the first input consists of truncated normalized fast Fourier transform (FFT) spectrograms [34], the second is constant Q transform (CQT) [7], and the third is FFT with a sliding window. Table II provides a brief summary of deep learning approaches for spoofing/genuine classification.

C. Discussion

Previous studies tackled the generalization problem by mainly investigating various feature selection and fusion ap-

proaches [31, 36, 37], studying feature representations and diverse classification approaches [30, 39, 42, 45], and applying diverse deep learning architectures such as ResNet [8, 30, 36] and RNN [9, 19, 31, 51, 52]. Furthermore, for deep learning approaches both raw audio data [9, 19, 30], and engineered features [8, 31, 36, 52] are considered. Nevertheless, neither the engineered features nor those learned automatically by deep learning succeeded in alleviating the generalization problem. In fact, there was a discrepancy between the training and testing performances; they improve the prediction for seen utterances, but they are not able to generalize to unseen ones. Nevertheless, the baseline approach [45], which is based on extracting the CQCC feature and GMM-based classification approach, outperforms the other state-of-the-art approaches in terms of generalization on the ASVspoof 2015 dataset but failed on the ASVspoof 2017 dataset. On the latter dataset, the approach proposed in [31] is the best reported approach with a testing EERor rate of 6.73%.

III. CLUSTERING

Clustering is an unsupervised learning approach that groups unlabeled instances into homogeneous clusters based on certain criteria or similar functions. This allowed the exploration and analysis of the data. There are three main approaches to clustering: (i) hierarchical clustering, (ii) partitioning, and (iii) density-based clustering approaches [5]. Hierarchical clustering [17] creates a hierarchy of clusters following either a top-down (divisive) or a bottom-up (agglomerative) strategy. Alternatively, partitioning or centroid-based clustering [32], is characterized by learning a representative of each cluster such as the cluster centers. Accordingly, data instances were assigned to the cluster corresponding to the closest representative. For this purpose, the distance to the representatives is calculated using distance metrics such as the Euclidean distance or Manhattan distance. On the other hand, density-based clustering approaches [28] consider the density rather than distance to assign an instance to a cluster. Clustering is performed in such a way that dense instances form clusters, whereas sparse instances are considered noise and outliers. Density-based approaches are characterized by the arbitrary shapes of the clusters. Clustering approaches can also be categorized as crisp or fuzzy. Although crisp clustering approaches assign an instance exclusively to one cluster, fuzzy clustering can assign an instance to more than one cluster using a membership degree. The latter can be perceived as an instance's probability of belonging to a given cluster. In this way, fuzzy approaches can deal with real-world applications in which clusters exhibit overlapping boundaries [5]. In the following section, we describe three fuzzy clustering approaches that will be investigated to uncover the underlying structure of voice-spoofing data. Specifically, we consider fuzzy c-means (FCM) clustering [4], simultaneous clustering and attribute discrimination (SCAD) [18], and competitive agglomeration CA [17] algorithms.

A. Fuzzy C-Means

Fuzzy c-means (FCM) [4] clustering performs a fuzzy partitioning of the unlabeled data by minimizing the intra-cluster distances. More precisely, for a set of instances, x_j , it simultaneously learns the cluster representatives (centers), c_i ,

and the fuzzy memberships, (u_{ij}) , by minimizing the objective function:

$$J(\mathbf{B}, \mathbf{U}; \mathcal{X}) = \sum_{i=1}^C \sum_{j=1}^N (u_{ij})^m \|x_j - c_i\|^2 \quad (1)$$

subject to

$$u_{ij} \in [0, 1] \& \forall i, j \text{ and } \sum_{i=1}^C u_{ij} = 1 \& \forall j \quad (2)$$

In Equation (1), x_j and $c_i \in R^d$ where d is the dimension of the vectors, m is a fuzzier that controls the membership fuzziness, C is the number of clusters and N is the number of instances.

B. Simultaneous Clustering and Attribute Discrimination

Simultaneous clustering and attribute discrimination (SCAD) [18] is an extension of FCM that addresses the problems of feature selection and aggregation. It learns the relevance feature weights, $\nu = [\nu_{ik}]_{i=1\dots c, j=1\dots d}$ with respect to each cluster. In addition to the centers, $C = [c_{ik}]_{i=1\dots c, j=1\dots d}$, and fuzzy membership, $U = [u_{ik}]_{i=1\dots c, j=1\dots N}$. This is achieved by minimizing the objective function, as follows:

$$J(\mathbf{C}, \mathbf{U}, \mathbf{V}; \mathcal{X}) = \sum_{i=1}^C \sum_{j=1}^N (u_{ij})^m \sum_{K=1}^d v_{ik} (x_{jk} - c_{ik})^2 \quad (3)$$

subject to

$$u_{ij} \in [0, 1] \& \forall i, j \text{ and } \sum_{i=1}^C u_{ij} = 1 \& \forall j \quad (4)$$

and

$$v_{ik} \in [0, 1] \& \forall i, k \text{ and } \sum_{i=1}^d v_{ik} = 1 \& \forall i \quad (5)$$

where C is the number of clusters, N is the number of instances and d is the feature size, and v_{ik} , c_{ik} and $u_{ij} \in R^d$ where d is the dimension of the vectors.

C. Competitive Agglomeration

Competitive agglomeration (CA) [17] is another extension of FCM that addresses the problem of estimating the number of clusters in an unsupervised manner. In fact, it learns the number of clusters while learning the cluster representatives and the fuzzy memberships. It combines hierarchical and partitioning clustering approaches, and thus benefits from their advantages. Specifically, CA applies the competitive agglomeration in order to select the best number of clusters. It begins by dividing the instances into small clusters. During the optimization process, the clusters compete over instances, and the empty clusters disappear gradually. The CA optimizes the following objective function:

$$J(\mathbf{B}, \mathbf{U}; \mathcal{X}) = \sum_{i=1}^C \sum_{j=1}^N (u_{ij})^2 \cdot d_{ij}^2(x_j, \beta_i) - \alpha \sum_{i=1}^C [\sum_{j=1}^N u_{ij}]^2 \quad (6)$$

where $B = (1, \dots, c)$ are the cluster representatives, $d_{ij}^2(x_j, \beta_i)$ is the distance between feature vectors x_j and prototype β_i , and u_{ij} is the fuzzy membership of instance j with respect to cluster i . As can be seen, the objective function in (6) incorporates two terms: the first one is inherited from the FCM objective function (1). On the other hand, the second term in (6) is the competitive term that allows cluster competition to enclose data instances.

IV. PROPOSED APPROACH

Owing to the high intra-class variance of the spoofing and genuine categories, the sub-groups of these two categories are scattered. Moreover, spoofing subgroups overlap with genuine subgroups, and vice versa. This renders the classification problem even more challenging. In fact, the learned classification model is too complex and may result in the overfitting of the training dataset. This reflects the low generalization of the supervised learning model. Therefore, we propose splitting each category into homogeneous groups, and then classifying unknown instances by considering the closest sub-group. The proposed spoofing countermeasure based on homogeneous subcategories is illustrated in Fig. 1. As one can see, it starts by extracting audio features from the genuine and spoofing utterances. Then, the genuine instances are clustered separately to determine the representatives of the genuine sub-categories in an unsupervised manner. Similarly, spoofing instances were clustered in order to obtain the spoofing representatives. The learned sub-category representatives are then used to classify unknown instances. Specifically, for the clustering task, we propose employing prototype-based fuzzy clustering approaches. This choice is motivated by the need to learn cluster representatives, and the fact that fuzzy memberships are better at handling the overlapping boundaries of clusters. In other words, we intend to investigate several prototype-based fuzzy clustering approaches such as FCM [4] and SCAD[17, 18]. In fact, FCM-based clustering approaches learn the cluster centers which is not the case for other types of clustering approaches such as density or hierarchical-based clustering algorithms. Moreover, SCAD learns the relevance feature weights while clustering the data. This allows for the automatic selection and aggregation of the features. Similarly, CA automatically estimated the number of clusters while clustering the data. The three clustering algorithms under consideration are optimized iteratively by alternating the update of the centers, the fuzzy memberships, and eventually the relevance feature weights and the number of clusters through the use of closed-form update equations. Furthermore, we plan to explore the number of clusters that generate the optimal subcategories for the spoofing and the genuine classes. In fact, estimating the number of clusters allows the correct structure of the data to be uncovered. Therefore, it yields a better local classification which helps to lessen the generalization issue for unseen instances. An illustrative example of the proposed spoofing countermeasure based on homogeneous subcategories is shown in Fig. 2. As it can be seen, the spoofing utterances are clustered into six clusters ($S1, S2, S3, S4, S5$, and $S6$), while genuine utterances are clustered into four clusters ($G1, G2, G3$, and $G4$). Then, the blue unseen instance was compared to the ten learned representatives before assigning it to one of the clusters. Because $G1$, one of the representatives of the genuine category, is the closest to the blue unseen instance, the latter

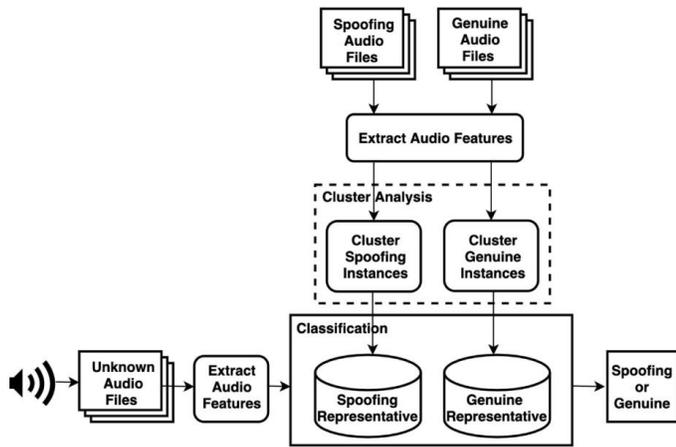


Fig. 1. General Overview of the Proposed Approach.

is classified as genuine.

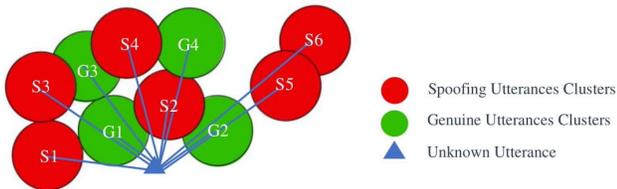


Fig. 2. Illustrative Example the Proposed Spoofing Countermeasure based on Homogenous Sub-Categories.

V. EXPERIMENTS

The dataset used for the experiments was ASVspoof 2017 version 2.0 [13]. The dataset contained audio files with a sampling rate of 16 kHz and a 16-bit resolution, which were divided into three subsets: a training set containing 3016 files, a development set containing 1710 files, and an evaluation set with 13,306 files. The training set had 1507 genuine and 1507 replay files, the development set had 760 genuine and 950 replay files, and the evaluation set had 1298 genuine and 12,008 replay files [13]. The Mel-frequency cepstral coefficients' (MFCCs) [11] and the constant Q cepstral coefficients' (CQCCs) [45] features are extracted from the audio files. MFCC is computed by applying the discrete cosine transform (DCT) type 2 on a 20 ms audio frame. This generates an audio spectrum that reflects energy in different frequency bands. After spectrum computation, a bank of triangular filters was employed to warp the spectrum into the Mel-scale. Finally, the results of a Mel-scale filter bank are logarithmized and decorrelated by applying the DCT. Alternatively, CQCC is based on a constant Q transform (CQT). The latter is a time-frequency analysis tool for short-time Fourier transform (STFT) [7]. The number of bins per octave was set to 12, and the sampling frequency was set to 44,000 Hz. The CQCCs' spectrum was derived by first performing CQT transform on the audio frame. Next, the logarithm non-linearity and linearization of the CQT's geometric scale were applied. Then, the final 167 CQCC cepstral coefficients were obtained by

applying the DCT [45]. Similar to the approaches described in Section III, the performance of the proposed approach is evaluated using an equal error rate (EER) [38]. It was calculated using a receiver operating characteristic (ROC) curve. More specifically, the EER is defined as the operating point where the false acceptance rate (FAR) and false rejection rate (FRR) are equal [16, 38].

A. Experiment 1: Discovering the Underlining Structure using Fuzzy C-Means

In this experiment, we clustered the genuine and the spoof classes from the training subset separately, using fuzzy c-means [4]. The same number of clusters was used for both classes, and it was tuned from 2 to 16 with a step of 2. The learned sub-category representatives were then used for the classification of unknown instances from the testing subset, using the K-nearest neighbor classifier KNN [3] with $K = 1$. The experiment was conducted on the CQCC, MFCC, and a concatenation of the CQCC and MFCC independently. The EER with respect to the number of clusters on the considered features is shown in Fig. 3.

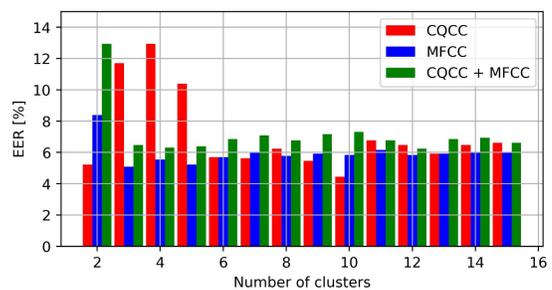


Fig. 3. EER with respect to the Number of Clusters when using Fuzzy C-Means [20] on CQCC, MFCC and the Concatenation of CQCC and MFCC.

As shown in Fig. 3, the EER varies with respect to the number of clusters and features considered. The best result is obtained when using a number of clusters equal to 2 for both classes on the CQCC feature. It reached an EER of 4.46%.

B. Experiment 2: Discovering the Underlining Structure using the Competitive Agglomeration

In this experiment, the underlining structure was learned using the competitive agglomeration (CA) [17] clustering algorithm in order to simultaneously cluster the training data and learn the optimal number of clusters. The same number of clusters was initially set to 100 for both classes (genuine and spoofed). Similar to Experiment 1, the learned cluster representatives are used for the classification of unknown instances from the testing subset, using the K-nearest neighbor classifier KNN [3] with $K = 1$. Moreover, an experiment was conducted on the CQCC, MFCC, and concatenation of the CQCC and MFCC independently. Table III reports the obtained EER, and the learned number of clusters with respect to the considered features.

As shown in Table III, the CQCC exhibited the lowest EER of 2.46%. Moreover, the optimal cluster learned by CA

TABLE III. EER AND THE LEARNED NUMBER OF CLUSTERS WHEN USING COMPETITIVE AGGLOMERATION (CA) [17] ON CQCC, MFCC, AND THE CONCATENATION OF CQCC AND MFCC

	CQCC	MFCC	CQCC+MFCC
EER	2.46%	14.86%	9.24%
No of genuine clusters	2	3	3
No of spoof clusters	2	2	3

TABLE IV. LEARNED FEATURE WEIGHTS WITH RESPECT TO EACH CLUSTER

	CQCC	MFCC
Cluster 1 (genuine)	0.92	0.08
Cluster 2 (genuine)	0.92	0.08
Cluster 1 (spoof)	0.91	0.09
Cluster 2 (spoof)	0.91	0.09

when using CQCC is two clusters for the genuine class and two clusters for the spoof class. This is similar to the result obtained in the first experiment using fuzzy c-means by tuning the number of clusters. We can then conclude that the CA clustering approach can learn the underlying structure of both the genuine and spoof classes while learning the optimal number of clusters.

C. Experiment 3: Discovering the Underlining Structure using Simultaneous Clustering and Attribute Discrimination

In this experiment, partitions of the genuine and spoof classes are mined using simultaneous clustering and attribute discrimination (SCAD) [18]. It aims to discover the underlying partitions while learning the optimal relevance feature weights of CQCC and MFCC audio features. In fact, the weights of each of these two considered features are learned with respect to each class. The number of clusters was set to two for the genuine class and two for the spoof class according to the results obtained in Experiment 2. The obtained partitions were then used for the classification task using the K-nearest neighbor classifier KNN [3] with $K = 1$. Table IV presents the obtained EER, and the learned number of clusters with respect to the considered features.

As reported in Table IV, the feature weights with the largest relevance were learned for the CQCC. This is in concordance with the results obtained in the previous experiments, which showed that the CQCC is more relevant for the classification of genuine/spoof utterances. To further investigate the CQCC feature, we applied SCAD to its 167 CQCC cepstral coefficients. In other words, each dimension of the CQCC is considered as a single feature. A feature relevance weight was then learned for each dimension. For Experiment 1, the number of clusters was tuned from 2 to 16 in steps of 2.

As shown in Fig. 4, when using the same number of clusters, the lowest EER, equal to 1.07, was obtained for the number of clusters equal to two. We can conclude then that performing SCAD on the CQCC yields better results. This is because it deals with the high dimension of CQCC by performing an optimal weighted sum of the coefficients. Fig. 5 shows the relevance feature weights for each cluster.

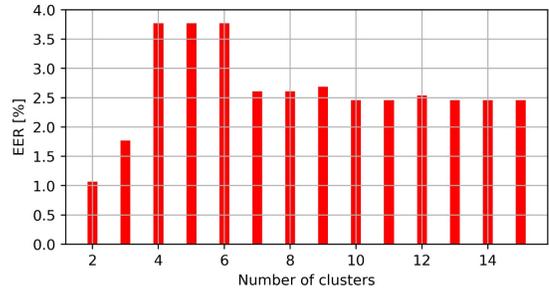


Fig. 4. EER with respect to the Number of Clusters when using SCAD [18] on CQCC Dimensions.

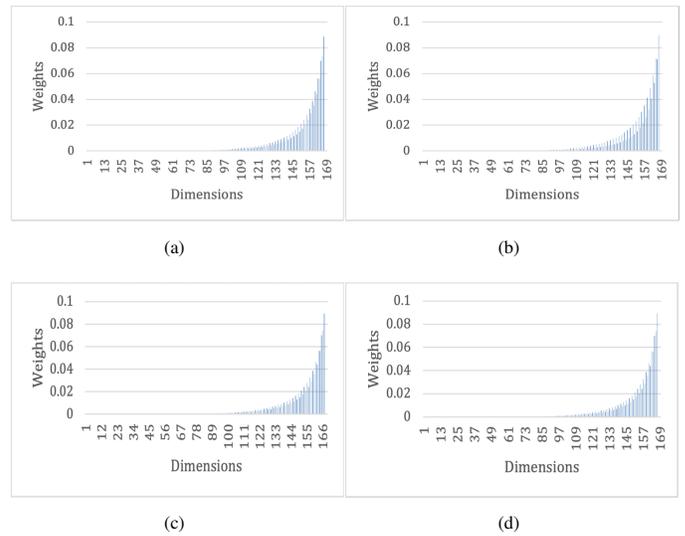


Fig. 5. Relevance Feature Weights with respect to (a) Cluster 1 (Genuine), (b) Cluster 2 (Genuine), (c) Cluster 1(Spoof), (d) Cluster 2 (Spoof).

D. Experiment 4: Performance Comparison with the State-of-the-Art Approaches

In this experiment, the performance of the proposed approach was compared to that of the state-of-the-art approach in the ASVspoof 2017. More specifically, the best results obtained using the considered clustering algorithms are compared to the conventional KNN [3] approach and to the countermeasures reported in the literature [8, 9, 27, 30, 31, 36]. To compare the proposed approach to KNN, we classified ASVspoof 2017 using KNN while tuning the neighboring parameter from 3 to 9. The experiment was conducted on MFCCs, CQCCs, features, and their corresponding concatenation. As shown in Fig. 6, the lowest EER (EER=3.62%) was obtained when using CQCC with K equal to 9.

Table V reports the training EER and the testing EER of the state-of-the-art approaches, the best KNN result, and the best results of the proposed approach with respect to the different clustering approaches under consideration.

As shown in Table V, the proposed approach outperforms the KNN classifier and the methods reported in the literature, regardless of the considered clustering algorithm. Moreover, it solves the generalization problem by reducing the performance

TABLE V. PERFORMANCE COMPARISON IN TERMS OF EER BETWEEN THE PROPOSED APPROACHES AND THE STATE-OF-THE-ART APPROACHES

Countermeasures	Training EER %	Testing EER%
Reported work in [30] (Features: Signal Logspec via FFT, Model: ResNet)	6.09	8.54
Reported work in [8] (Features: CQCC and MFCC, Model:GMM + ResNet)	2.58	13.30
Reported work in [36](Features: Fusion of HFCC and CQCC, Model:DNN + SVM)	7.6	11.5
Reported work in [9](Features: MFCC, Fbank, Model: LSTM + GRU RNN)	6.32	9.81
Reported work in [31](Features: CQT and FFT, Model: LCNN, SVM, CNN + RNN)	3.95	6.73
Reported work in [27](Feature: CQCC, Model: GMM)	10.35	24.77
KNN (K=9, Feature: CQCC)	1.05	3.62
Proposed approach based on Fuzzy C-Means (No of genuine clusters= 2, No of spoof clusters=2, Feature: CQCC)	3.15	4.46
Proposed approach based on CA (feature: CQCC)	2.67	2.46
Proposed approach based on SCAD No of genuine clusters= 2, No of spoof clusters=2, Features: coefficient of CQCC)	0.13	1.07

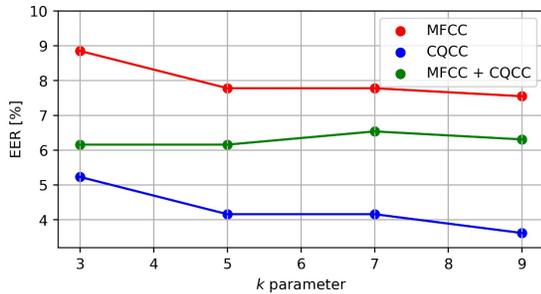


Fig. 6. EER with respect to the Number of Neighbor Parameter K when using KNN [3] on CQCC, MFCC, and the Concatenation of CQCC and MFCC.

gap between the training EER and the testing EER. This is achieved by mining the underlying structure of both genuine and spoof utterances. Furthermore, the lowest EER is obtained when using SCAD with a number of genuine clusters equal to 2, and a number of spoof clusters equal to 2 on CQCC coefficients. The achieved EER was 1.07%. This improved the result by a ratio of 5.66% compared with the best reported work [31], which achieved an EER of 6.73%.

VI. CONCLUSION

Voice spoofing is a prominent security risk that requires effective countermeasures to protect the user’s information when using ASV systems. These countermeasures amount to the classification of the utterances into genuine or spoofing categories. However, this classification problem is challenging because of high class variance. In fact, genuine utterances are subject to variability due to the differences in speakers’ voices and the discrepancies within the human voice due to emotions or other effects. Similarly, spoofing utterances are subjected to the same variability, in addition to the variability caused by the recording devices employed. Moreover, the diversity in the methods that produce spoofing utterances, such as voice manipulation and synthesis, contributes significantly to the variance in the spoofing class. This high variance in utterances drastically affects the performance of the spoofing classification task. Specifically, it limits the model’s generalization and yields a less accurate system. Recently, considerable attempts have been made to address the low generalization of spoofing countermeasures. The reported works focused mainly on investigating various feature selection and fusion approaches, studying feature representations, and applying diverse deep learning architectures. However, neither handcrafted features nor deep learning-based descriptors have succeeded in alleviating the

generalization problem. In fact, although they have shown slight improvements in the prediction performance of the models, they fail to generalize to unseen utterances. In fact, they are prone to overfitting, as indicated by the discrepancy between the training and testing performances. In this study, we devised a new countermeasure to address the low-generalization problem. Specifically, the proposed approaches mined the understructure of the genuine and spoofing utterances. This was achieved by integrating the clustering component into the classification process. The experimental results showed that mining hidden partitions of voice utterances using fuzzy clustering yielded a better generalization of the voice-spoofing countermeasure. In fact, the proposed approach outperformed the state-of-the-art approaches. Specifically, when using the CA clustering approach, the training and testing EERs were similar. Moreover, when using SCAD on the CQCC feature for a number of genuine clusters equals to 2, and a number of spoof clusters equal to 2, the performance is drastically improved with an EER of 1.07%. In future work, we suggest investigating additional audio features. Moreover, we intend to use CA as the first step in order to learn the number of clusters and the initial fuzzy memberships. Then, SCAD would be performed using the obtained results.

ACKNOWLEDGMENT

This work was supported by the Research Center of the College of Computer and Information Sciences at King Saud University. The authors are grateful for this support.

REFERENCES

- [1] Moez Ajili. Reliability of voice comparison for forensic applications. Avignon, 2018.
- [2] Federico Alegre, Asmaa Amehraye, and Nicholas Evans. A one-class classification approach to generalised speaker verification spoofing countermeasures using local binary patterns. In *IEEE 6th International Conference on Biometrics: Theory, Applications and Systems, BTAS 2013*, pages 1–8, October 2013.
- [3] N. S. Altman. An introduction to kernel and nearest-neighbor nonparametric regression. *American Statistician*, 46(3):175–185, 1992.
- [4] James C. Bezdek. *Pattern Recognition with Fuzzy Objective Function Algorithms*. Kluwer Academic Publishers, USA, 1981.
- [5] Christopher M Bishop and Nasser M Nasrabadi. *Pattern recognition and machine learning*, volume 4. Springer, 2006.
- [6] Bernhard Boser, Isabelle Guyon, and Vladimir Vapnik. A Training Algorithm for Optimal Margin Classifier. *Proceedings of the Fifth Annual ACM Workshop on Computational Learning Theory*, 5, Aug 1996.
- [7] Judith C. Brown. Calculation of a constant q spectral transform. *The Journal of the Acoustical Society of America*, 89(1):425–434, 1991.
- [8] Zhuxin Chen, Zhifeng Xie, Weibin Zhang, and Xiangmin Xu. ResNet and model fusion for automatic spoofing detection. In *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, pages 102–106, 2017.

- [9] Zhuxin Chen, Weibin Zhang, Zhifeng Xie, Xiangmin Xu, and Dongpeng Chen. Recurrent neural networks for automatic replay spoofing attack detection. In *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, pages 2052–2056, 2018.
- [10] Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Fethi Bougares, Holger Schwenk, and Y Bengio. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. Jun 2014.
- [11] Steven B. Davis and Paul Mermelstein. Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 28(4):357–366, 1980.
- [12] Najim Dehak, Réda Dehak, Patrick Kenny, Niko Brummer, Pierre Ouellet, and Pierre Dumouchel. Support vector machines versus fast scoring in the low-dimensional total variability space for speaker verification. In *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2009.
- [13] Héctor Delgado, Massimiliano Todisco, Md Sahidullah, Nicholas Evans, Tomi Kinnunen, Kong Aik Lee, and Junichi Yamagishi. ASVspoof 2017 Version 2.0: meta-data analysis and baseline enhancements. 2018.
- [14] P Ding and L Zhang. Speaker Recognition Using Principal Component Analysis. *Proc. ICONIP2001*, Jan 2001.
- [15] Project Editor. DRAFT INTERNATIONAL STANDARD ISO / IEC DIS 30107-3 Information technology — Biometric presentation attack detection. 2017.
- [16] Tom Fawcett. An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8), 2006.
- [17] Hichem Frigui and Raghu Krishnapuram. A robust competitive clustering algorithm with applications in computer vision. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21(5):450–465, 1999.
- [18] Hichem Frigui and Olfa Nasraoui. In *Ninth IEEE International Conference on Fuzzy Systems. FUZZ-IEEE 2000 (Cat. No. 00CH37063)*, volume 1, pages 158–163. IEEE, 2000.
- [19] Alejandro Gomez-Alanis, Antonio M. Peinado, Jose A. Gonzalez, and Angel M. Gomez. A deep identity representation for noise robust spoofing detection. In *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, pages 676–680, 2018.
- [20] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT press, 2016.
- [21] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *European conference on computer vision*, pages 630–645. Springer, 2016.
- [22] Guy A. Hembury, Victor V. Borovkov, Juha M. Lintuluoto, and Yoshihisa Inoue. Deep Residual Learning for Image Recognition Kaiming. *Chemistry Letters*, 32(5):428–429, 2003.
- [23] H.-G Hirsch and D Pearce. The AURORA experimental framework for the performance evaluation of speech recognition systems under noisy conditions. *Proceedings of the 6th International Conference on Spoken Language Processing (ICSLP)*, 4:29–32, Jan 2000.
- [24] Sepp Hochreiter and Jürgen Schmidhuber. Long Short-term Memory. *Neural computation*, 9(8):1735–1780, Dec 1997.
- [25] Alan Izenman. *Linear Discriminant Analysis*. Springer, New York, NY, Jan 2008.
- [26] Madhu R. Kamble, Hardik B. Sailor, Hemant A. Patil, and Haizhou Li. Advances in anti-spoofing: From the perspective of ASVspoof challenges. *APSIPA Transactions on Signal and Information Processing*, 9, 2020.
- [27] Tomi Kinnunen, Md Sahidullah, Héctor Delgado, Massimiliano Todisco, Nicholas Evans, Junichi Yamagishi, and Kong Aik Lee. The ASVspoof 2017 challenge: Assessing the limits of replay spoofing attack detection. In *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2017.
- [28] Hans-Peter Kriegel, Peer Kröger, Joerg Sander, and Arthur Zimek. Density-based Clustering. *Wiley Interdisc. Rev.: Data Mining and Knowledge Discovery*, 1(3):231–240, May 2011.
- [29] Yoohwan Kwon, Soo Whan Chung, and Hong Goo Kang. Intra-class variation reduction of speaker representation in disentanglement framework. In *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, pages 14–18, Shanghai, China, September 2020.
- [30] Cheng-I Lai, Alberto Abad, Korin Richmond, Junichi Yamagishi, Najim Dehak, and Simon King. Attentive Filtering Networks for Audio Replay Attack Detection. In *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, pages 6316–6320. IEEE, 2019.
- [31] Galina Lavrentyeva, Sergey Novoselov, Egor Malykh, Alexander Kozlov, Oleg Kudashev, and Vadim Shchemelinin. Audio replay attack detection with deep learning frameworks. In *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, pages 82–86, 2017.
- [32] Y Leung, J Zhang, and Z Xu. Clustering by space-space filtering. *IEEE Trans. Pattern Anal. Mach. Intell.*, 22(12):1396–1410, Jan 2000.
- [33] Qi Li. In *2009 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pages 181–184. IEEE, 2009.
- [34] Charles Loan. *Computational Frameworks for the Fast Fourier Transform*. SIAM, Jan 1992.
- [35] L R Medsker and L C Jain. *Recurrent Neural Networks: Design and Applications*. International Series on Computational Intelligence. CRC Press, 1999.
- [36] Parav Nagarsheth, Elie Khoury, Kailash Patil, and Matt Garland. Replay attack detection using DNN for channel discrimination. In *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, pages 97–101, 2017.
- [37] Sergey Novoselov, Alexander Kozlov, Galina Lavrentyeva, Konstantin Simonchik, and Vadim Shchemelinin. STC anti-spoofing systems for the ASVspoof 2015 challenge. pages 5475–5479. Mar 2016.
- [38] John Oglesby. What’s in a number? Moving beyond the equal error rate. *Speech Communication*, 17(1-2), 1995.
- [39] Tanvina B. Patel and Hemant A. Patil. Combining evidences from mel cepstral, cochlear filter cepstral and instantaneous frequency features for detection of natural vs. spoofed speech. In *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, Germany, September 2015.
- [40] T.F. Quatieri. *Discrete-time Speech Signal Processing: Principles and Practice*. Pearson Education Taiwan, 2005.
- [41] Md Sahidullah, Héctor Delgado, Massimiliano Todisco, Tomi Kinnunen, Nicholas Evans, Junichi Yamagishi, and Kong Aik Lee. Introduction to voice presentation attack detection and recent advances. In *Handbook of biometric anti-spoofing*, pages 321–361. Springer, 2019.
- [42] Md Sahidullah, Tomi Kinnunen, and Cemal Haniilçi. A comparison of features for synthetic speech detection. In *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2015.
- [43] Zia Saquib, Nirmala Salam, Rekha Nair, and Nipun Pandey. Voiceprint Recognition Systems for Remote Authentication-A Survey. *International Journal of Hybrid Information Technology*, 4(2), April 2011.
- [44] Simone Scardapane, Lucas Stoffl, Florian Rohrbain, and Aurelio Uncini. On the use of deep recurrent neural networks for detecting audio spoofing attacks. In *Proceedings of the International Joint Conference on Neural Networks*, pages 3483–3490. IEEE, 2017.
- [45] Massimiliano Todisco, Héctor Delgado, and Nicholas Evans. A New Feature for Automatic Speaker Verification Anti-Spoofing: Constant Q Cepstral Coefficients. In *Odyssey 2016*, pages 283–290, 2016.
- [46] Xiang Wu, Ran He, Zhenan Sun, and Tieniu Tan. A light CNN for deep face representation with noisy labels. *IEEE Transactions on Information Forensics and Security*, 13(11):2884–2896, 2018.
- [47] Zhizheng Wu, Eng Chng, and Haizhou Li. Detecting Converted Speech and Natural Speech for anti-Spoofing Attack in Speaker Recognition. *13th Annual Conference of the International Speech Communication Association 2012, INTERSPEECH 2012*, 2, Jan 2012.
- [48] Zhizheng Wu, Nicholas Evans, Tomi Kinnunen, Junichi Yamagishi, Federico Alegre, and Haizhou Li. Spoofing and countermeasures for speaker verification: A survey. *speech communication*, 66:130–153, 2015.
- [49] Zhizheng Wu, Tomi Kinnunen, Nicholas Evans, Junichi Yamagishi, Cemal Haniilçi, Md Sahidullah, and Aleksandr Sizov. ASVspoof 2015: The first automatic speaker verification spoofing and countermeasures challenge. In *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2015.
- [50] Yong Xu, Qiuqiang Kong, Qiang Huang, Wenwu Wang, and Mark D. Plumbley. Attention and localization based on a deep convolutional recurrent model for weakly supervised audio tagging. In *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2017.
- [51] Sung-Hyun Yoon, Hee-Soo Heo, Hye-Jin Shim, Ha-Jin Yu, and Jee-Weon Jung. Replay Spoofing Detection System for Automatic Speaker Verification Using Multi-Task Learning of Noise Classes. pages 172–176, 2018.
- [52] Chunlei Zhang, Chengzhu Yu, and John H.L. Hansen. An investigation of deep-learning frameworks for speaker verification antispoofing. *IEEE*

- Journal on Selected Topics in Signal Processing*, 11(4):684–694, 2017.
- [53] Xinhui Zhou, Daniel Garcia-Romero, Ramani Duraiswami, Carol Espy-Wilson, and Shihab Shamma. Linear versus mel frequency cepstral coefficients for speaker recognition. In *2011 IEEE Workshop on Automatic Speech Recognition Understanding*, pages 559–564. IEEE, Dec 2011.

PhishRepo: A Seamless Collection of Phishing Data to Fill a Research Gap in the Phishing Domain

Subhash Ariyadasa¹, Shantha Fernando² and Subha Fernando³

Department of Computational Mathematics, University of Moratuwa, Sri Lanka^{1,3}

Department of Computer Science and Engineering, University of Moratuwa, Sri Lanka²

Department of Computer Science and Informatics, Uva Wellassa University, Sri Lanka¹

Abstract—Machine learning-based anti-phishing solutions face various challenges in collecting diverse multi-modal phishing data. As a result, most previous works have trained with little or no multi-modal data, which opens several drawbacks. Therefore, this study aims to develop a phishing data repository to meet the diverse data needs of the anti-phishing domain. As a result, a gap-filling solution named PhishRepo was proposed as an online data repository that collects, verifies, disseminates, and archives phishing data. It includes innovative design aspects such as automated submission, deduplication filtering, automated verification, crowdsourcing-based human interaction, an objection reporting window, and target attack prevention techniques. Moreover, the deduplication filter, used for the first time in phishing data collection, significantly impacted the collection process. It eliminated the duplicate data, which causes one of the most common machine learning errors known as data leakage. In addition, PhishRepo enables researchers to apply modern machine learning techniques effectively and supports them by eliminating phishing data hassle. Therefore, more thoughtful use of PhishRepo will lead to effective anti-phishing solutions in the future, minimising the social engineering crime called phishing.

Keywords—Cyberattack; crowdsourcing; internet security; phishing; machine learning; multi-modal data

I. INTRODUCTION

Industry 4.0, or the fourth industrial revolution that marks the beginning of the imagination age, has opened various opportunities for human beings through automation and data exchange. However, it is a double-edged sword where criminals also optimise the revolution change to effectively operate their criminal activities on the Internet. Phishing is an illegal activity that relies on the Internet, which has gained a top rank in the cyber threat landscape [1]. It is a social engineering threat that damages Internet users illegally using digital assets—incidentally personal and confidential information [2]. Phishing is known as ‘identity theft’ because it impersonates one’s identity in cyberspace for the phisher’s benefit [3], [2].

The phishing threat first occurred in 1996 [2], and initially, online banking and e-commerce services were popular among phishers [4]. The direct or indirect financial gains motivate phishers in phishing, and fame and notoriety are also attractive [3]. Phishers are constantly moving with technology. Therefore, they are keen to experiment and improve attacking strategies in the phishing domain without failing in front of the available security countermeasures [5], [6]. The number of phishing attacks is still rising. Interestingly, the Anti-Phishing Working Group (APWG) has stated that phishing attacks had doubled in 2020, and in October 2020, only they have detected 225,304 unique phishing websites [7].

In phishing attacks, phishers commonly send an email to a user with an embedded link to redirect the user to a phishing site [5], [2]. This email often denotes a specific scenario like updating account details or security upgrades and creates a way to convince users to believe it [2]. The phishers recently used the Coronavirus pandemic (COVID-19) to raise phishing campaigns to fool Internet users [2]. However, by accepting, an unsuspecting user might click on the given link and move to the phishing website, which is very similar to the legitimate website, to feel more confident about his previous action [5], [2]. Then, the most dangerous thing happens in the process. By believing this is the legitimate website, the unsuspecting user enters his vital information to the phisher’s website that impersonates the legitimate website. That information could be bank details, login credentials, a social security number, a credit card number, or other personal or confidential information [2]. However, this would be the main harvest of the phisher of this phishing process, and he might use it or sell it for his benefit.

According to the literature, 95% of phishing attacks were succeeded due to human errors [2]. Therefore, numerous solutions [8], [9], [10], [11], [12], [13], [14], [15], [16], [17], [18], [19] were introduced in the last two decades to safeguard humans from this prevalent Internet threat by detecting phishing attacks. Those solutions could be mainly categorised into user education and software-based solutions [20]. Out of those categories, software-based solutions were more prominent in the past since user education is associated with a high cost and requires fundamental knowledge of computer security [2]. The software-based solutions also use different approaches when finding an effective anti-phishing solution [20]. Of those approaches, machine learning shows promising results due to its unique advantages like handling frequent data changes and automating the learning process [6]. However, machine learning studies in the phishing domain suffer from labelled phishing data [21], [22]. Therefore, researchers primarily work with their data [12], [15], [17], [16], [23] due to a lack of benchmark datasets available in the current anti-phishing domain and their limitations [22].

The current study mainly focuses on finding an effective solution to the difficulty of getting labelled data or training with a limited amount of features in the phishing domain. It is essential in the present context since advanced techniques like deep learning can work effectively with many high-dimensional data when identifying complex attacks like phishing [24]. Further, these labelled data may be more effective in retraining the trained machine learning models since phishing

attacks are rapidly changing over time [21], [22]. The proposed solution to fill the identified research gap in the phishing domain is an online repository that can constantly collect, verify, disseminate, and archive real-time phishing data. This solution allows automatic submission of phishing data by anti-phishing solutions and guarantees the diversity of data through different filters like deduplication.

Further, it effectively uses existing phishing verification systems and crowdsourcing techniques to review the labelling of those collected data. Moreover, it manages the essential aspects of the submitted phishing records to open data to the scientific community, especially for anti-phishers. The main contributions of this study are an online phishing data repository for collecting, validating, disseminating and archiving real-time phishing data; a large-scale, diverse phishing data in raw format for research purposes; and a set of design artefacts to have in a real-time phishing data repository.

This article aims to introduce the gap-filling solution that touches the data needs in the anti-phishing domain and demonstrates the effectiveness of the used architecture of PhishRepo in the problem domain. The other sections of this paper include Section II - a high-level description of the problem domain; Section III - a review of the literature; Section IV - the architecture of the proposed solution; Section V - experiments were performed on the collected data to demonstrate the diversity and effectiveness of the machine learning process; Section VI - a discussion of the significance and usage of PhishRepo, and finally, Section VII - the concluding remarks of the study.

II. PROBLEM DEFINITION

Machine learning-based anti-phishing solutions mainly have two steps [17]. First, the features required to detect phishing attacks are extracted and then a machine learning model is trained using the extracted features. The feature extraction happens based on multiple information sources available on a website. The Uniform Resource Locator (URL) of a website is the popular source for many of the recent anti-phishing solutions [12], [15], [25], [23], and third party-based features like Alexa ranking and age of the domain are also used in different solutions [26], [27], [28], [29]. The website content, either human-readable or markup content, refers to HTML content, another vital source for extracting features. It has been used in many recent studies [14], [16], [17], [18], [19] to extract different features for the learning model. Further, several studies [13] already used captured images of the web page (i.e., a screenshot) when training machine learning-based anti-phishing solutions.

Supervised learning is the dominant practice with many existing anti-phishing solutions in the machine learning area [12], [14], [15], [17]. Therefore, the labelled data is essential for training the learning model. However, constructing a large-scale, diverse phishing dataset effective in training is impossible in one night since the phishing websites are short-lived [30], [31]. Therefore, it should be a continuous process and take time. However, phishing verification systems such as PhishTank (<https://phishtank.org/>) and OpenPhish (<https://openphish.com/>) collect many phishing URLs [30], including optional information like the screenshot of the website page

and network information (i.e., WHOIS information). Although these systems contain the URLs, those do not include all information sources required to extract the most recently exercised feature vectors directly [22]. It is a downside of these verification systems, and it negatively impacts research since the researchers need a systematic way to collect data in the initial step of their methodology. Since the data collection takes much time, many machine learning-based anti-phishing studies used less data during the training phase [28], [32], [33]. For example, [32] used 1,428 phishing data, and [28] used 2,000 phishing data during their experiments. However, some of the accessed solutions that used more data in model training are URL based solutions, and those may not be effective due to the challenges that exist only with URL based information [21]. Therefore, multi-modal features, marked as effective in phishing detection due to the representation of many phishing attack characteristics [34], are essential in the present phishing detection context. Free public access to such data sources is necessary for better detection solutions in future.

Moreover, the literature has shown that the researchers use old datasets due to the lack of new public datasets [22]. It results in inept learning models on recent phishing attacks [22]. These factors highlight the importance of an organised way of acquiring the latest multi-modal feature enabled diverse phishing data for future phishing detection. Therefore, the difficulty of getting labelled data or training with a limited number of features or data has become a significant problem in the machine learning-based phishing detection area that should be resolved to expect promising results in future research [21]. This study will resolve the identified issues by answering two questions: how can a phishing data repository be made to support anti-phishing research effectively? and what are the most effective design strategies that could be used in a real-time phishing data repository to collect, verify and disseminate large-scale, diverse phishing data?

III. RELATED LITERATURE

As highlighted in the problem definition, the difficulty of getting the latest, labelled phishing data with multi-modal features is a significant research challenge in machine learning-based phishing detection. This challenge could be overviewed closely by the three most related topics to the current study: data collection for phishing websites detection, feature selection for phishing websites detection, and data labelling in machine learning.

A. Data Collection for Phishing Websites Detection

Phishing techniques are constantly evolving due to technological improvements, enhanced security countermeasures and educated public [2]. Early days, phishers used untargeted attacks, and unsuspecting users were caught [2]. However, now phishers are more into target attacks, and techniques like spear-phishing are more prominent in the phishing domain [2]. The literature highlighted that the success rate of untargeted phishing attacks is less than 5%, while 19% of target attacks like spear-phishing get success [6]. However, when the phishing threat grew, many different parties like brand owners, researchers, and law enforcement were interested in these attacks from different perspectives [35]. Therefore, numerous organisations like APWG, Phishing Incident Response Team,

Phishing Report Network, and Digital PhishNet started to collect phishing attack-related data, resulting in different levels of data collection [36]. Further, an organisation like APWG mainly depends on the public, anti-phishing working groups, Internet service providers and brand owners when collecting phishing data [36].

In contrast, the Phisherman project [36], [35] addressed this phishing data collection process differently. It changed the present way of collecting phishing data and introduced a web-based system to collect, validate, disseminate and archive real-time phishing data. It is a global data collection system and fulfils three basic requirements: submitting suspicion records, saving the records for future use and outputting historical phishing data to interested parties. Phisherman has used an automated phishing records verification process, and the submitted records are verified in two steps. However, the first step is only for the submissions collected from individuals and high-volume spam feeds. The important feature in Phisherman in the current study's perspective is the dissemination of the collected data. Phisherman supports the data distribution in two ways: subscription and queries. However, these options allow downloading a blocklist or a full incident report in XML format.

PhishTank is one of the favourites to collect phishing data by many anti-phishing tool introducers [15], [34], [33], [23]. PhishTank was launched in 2006, and it is a community-based phishing verification system [30]. The PhishTank facilitates submitting phishing URLs, and the community votes those submissions to be a phishing website or marked as a legitimate website. However, when looking at those studies, the researchers used the PhishTank only to get phishing URLs and have not been used to extract other information sources like the screenshot of the phishing website and WHOIS information present in some of the submitted phishing records. The data distribution strategy used by the PhishTank might be the closest reason for such a trend.

OpenPhish is another phishing verification system that collects phishing data via an autonomous phishing detection algorithm shaped through research. It is also popular among anti-phishers [30], and it distributes the collected data like URL, target brand and screenshot to interested parties. However, OpenPhish is not for free, and a free account gets only the phishing URLs, which also gets in every twelve-hour frequency.

The UCI Phishing dataset available in the University of California - Irvine's (UCI) repository is also popular among researchers in the phishing domain [31]. However, it is an old dataset with limited data (i.e., 11,055 maximum). Further, it has a preprocessed set of features and bounds the research scope to those features. It is one of the main drawbacks of this dataset, and the heuristics used to preprocess those data [37] are also not examined in the current environment. Similarly, [22] presented high quality, a diverse phishing data source for benchmarking purposes, and one output of their study is implementing a benchmarking framework called PhishBench. The dataset was constructed using the sources like PhishTank, OpenPhish, and APWG and a systematic approach was undertaken when collecting data. However, it needs such an approach again when collecting new data, which may be costly and time taken.

Furthermore, Web2Vec [19] and PhishPedia [38] are another two datasets collected with the support of PhishTank and OpenPhish. PhishPedia collected phishing records from OpenPhish using a premium account to get additional information like target brands. However, these datasets also contain old data compared to today and since these studies are not focused on updating these datasets, implementing anti-phishing tools to detect the latest phishing attacks is problematic.

Phisherman project is the only landmark for a deliberate phishing data collection and dissemination approach. However, Phisherman is not publicly available [39]. Therefore, it is not a solution for the identified data collection problem. The solutions like PhishTank and OpenPhish have different intentions, such as maintaining blocklists and identifying target brands. Those data collections are more into URL related information extraction in phishing detection, therefore not effective in the data collection problem mentioned in this study. The individual data sources are an excellent approach to donating phishing data to others; however, the relevancy depends on the frequent update and the ability to support multi-modal features in the present machine learning-based anti-phishing domain.

B. Feature Selection for Phishing Websites Detection

Feature selection is essential in phishing detection research since it impacts detection accuracy [6], [22]. The researchers in the literature introduce different feature sets that represent the essential information sources that need to include in a dataset. A more complex categorisation of phishing features is found in [22]. They used more than 250 phishing detection-based studies and divided the phishing features mainly into two classes: URL and website. Again, those two classes divide into lexical, network and script level features. Then these features were furthermore analysed based on the format and categorised into three. 1). Syntactic - syntactic correctness (i.e., port number and Term frequency-inverse document frequency referred to as TF-IDF), 2). Semantic - the meaning and interpretation of the content (i.e., presence of the target brand in URL and web page), and 3). Pragmatic - the features do not directly relate to syntax or meaning (i.e., backlisted words in a URL, WHOIS information, and script loading time).

In a similar study, phishing features were primarily categorised into four feature sets [6]. 1). URL-based lexical features like the length of the URL and the presence of the HTTPS protocol, 2). URL-based host features like WHOIS information, 3). web page content features like page rank, hyperlinks and forms in the HTML content, and 4). visual similarity-based elements like images and colours. Further, [14] used only URL and web page content features in their study, and HTMLPhish [17] is a particular case that used superior web page content features in phishing detection. Furthermore, [40] introduced another set of features in their research on machine learning-based phishing attacks. In that, they have mentioned four main groups of features, namely, URL-based features (i.e., number of subdomains, length, and number of digits), domain-based features (i.e., age of the domain, and whether it is blocked in reputed services), page-based features (i.e., page rank, and Alexa rank) and content-based features (i.e., page title, body text, a web page screenshot, and images).

After analysing the available feature sets in explored literature, it is clear that the URL and the web page are the most

important information sources to extract different features for model training. Although a website could be callable if the URL is saved in general, the phishing web pages cannot be recovered only from the URL since those are short-lived [30], [31]. Therefore, the instant saving of the phishing page and relevant resources like the web page screenshot, images, CSS, and JavaScript when the attack is active and online is essential for future use [22].

Moreover, the screenshot of the web page is an essential feature to consider in the machine learning-based visual similarity area [13]. Therefore, the study identified three primary sources for feature extraction for machine learning-based phishing detection. These are URL, web page, and third-party services. Further, those information sources are essential to consider when constructing an adequate dataset for future research since it supports the extraction of all the required features from one dataset.

C. Data Labelling in Machine Learning

Machine learning-based anti-phishing solutions are more toward the supervised learning paradigm. Therefore, labelled data is essential for model training [21], and expert labelling is a popular approach when labelling data in machine learning [41]. However, expert-based labelling is often costly and time-consuming since modern machine learning needs large-scale datasets [41]. Due to the limitations of the expert approach, crowdsourcing has become a widespread technique in data collection [42], [41], [43]. Crowdsourcing is based on collective intelligence, which beliefs together is better than a single entity [44]. It has advantages like low cost, fast labelling and diverse opinions than the expert approach [43]. However, the main drawback is getting high-quality labelled data [41], [43]. Factors like the poor commitment of workers, uncertainty in the tasks, prior knowledge of the given task, and novice workers are some of the reasons for imperfect quality labels in crowdsource based data labelling [41].

However, the quality of labels in crowdsourcing could be improved through several techniques discussed in the literature. Those are pre-training [41], task pricing [41], [43], calculation of labelling quality of workers [45], [46], and peer review of the crowd worker work [47]. As mentioned in [46], identifying incorrect labelling data points is not sufficient in crowdsourcing since the labelling quality of the workers also matters. In another study, [42] proposed a relabeling strategy called absolute cumulative majority relabeling (ACMR). It allows relabeling of the same data point multiple times. It uses a voting mechanism to select majority voting, and if a label achieves more than 50% voting, it sets that as the correct label for that data point. However, if none of the labels could earn more than 50% are discarded in the ACMR strategy. Revolt [41] is another solution that uses crowdsourcing when collecting data for machine learning tasks. In revolt, the dataset is divided into multiple batches for the crowd workers' easiness, and groups collectively contribute to each batch. From a different perspective, [48] studied the effect of cognitive biases in crowdsourcing. The study identified that the anchoring, bandwagon, and decoy effects occur in crowdsourcing, and an experiment has shown that a 28% accuracy loss was recorded due to the anchoring effect.

As identified, crowdsourcing is a practical approach to phishing data labelling. Further, the quality of the workers needs to be evaluated, and peer review of the workers' work is essential to maintain the quality. ACMR is a better strategy for phishing labelling since it allows adding many labels to a single record, and a majority voting technique is used when selecting an appropriate label. Multiple batches in the labelling process are also aligned with the current study since it reduces the overhead of seeing more labelling tasks simultaneously. Further, avoiding the dependency on one information or specific people is vital in crowdsourcing-based labelling and getting a quick explanation about the submitted label, as Revolt [41] proposed, is essential in current work to avoid doubtful labels. However, the task pricing and pre-training are not applicable here since the cost is incurred with those techniques, and the proposed solution is freeware.

IV. PHISHREPO

PhishRepo is an online phishing data repository for collecting, validating, disseminating and archiving real-time phishing data. The researchers and other interested parties would use it for their customised needs associated with data. PhishRepo architecture consists of three primary modules: input, verification, and distribution.

PhishRepo is a data collection solution for industry 4.0. Therefore, it introduces a safe architecture to integrate with autonomous anti-phishing tools to submit their phishing data directly to the repository for others' use. Therefore, this solution can provide effective results by collaborating with existing anti-phishing solutions. The other speciality of the repository is the type of data it collects. The repository is designed to collect most information sources, namely URLs, web pages and third-party service information relevant to a submission. However, when collecting third-party service information, the repository limits the free account level and possible third-party details are stored in the phishing records. Following is an in-depth explanation of PhishRepo modules.

A. Input Module

The primary task of the input module is the collection of phishing data which includes collecting phishing URLs from outside, acquiring the relevant information sources and archiving those in the repository, as shown in Fig. 1. Phishing data collection seems challenging since 63% of the phishing campaigns last within the first two hours [20]. It is not challenging to archive only the URLs [22]. However, PhishRepo is responsible for collecting URLs, web pages and possible third-party information sources for each inputted phishing record. Therefore, the detection time and reporting time are crucial in data collection. Thus, as a specific design consideration, PhishRepo allows external anti-phishing tools to submit their findings (i.e., phishing URLs) automatically. Then, it minimises the difference between detection and reporting time. It also helps collect the most active and online phishing URLs to effectively acquire the required information sources. Although the manual submission exists in PhishRepo, as in Fig. 1, PhishRepo encourages automated submissions to get the most active and online phishing URLs in the data collection process.

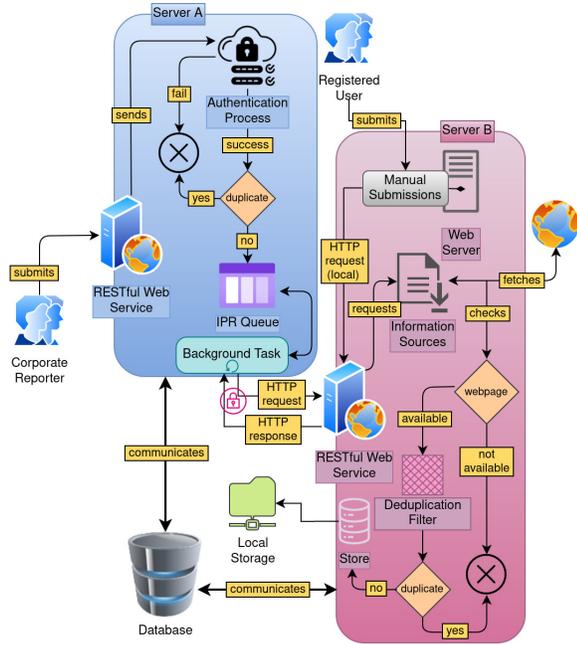


Fig. 1. Workflow Diagram of the Input Module

PhishRepo’s input module consists of five components: authentication, accumulation, deduplication, targeted attack prevention (TAP), and manual submission. The followings discuss these components in detail.

1) *Authentication*: PhishRepo has five users: administrator, editor, reporter, beneficiary, and guest. The administrator is the primary account holder with full privileges, and the editor is responsible for verifying the submitted phishing records. There are three levels in the editor account: newbie, competent, and expert. These levels are achieved by each user based on their performance. However, the expert editor is a chief editor type in PhishRepo and is responsible for the final decision of incorrect submission. The expert editor can modify the records if required and report phishing instantly. Therefore, the automatic account upgrade is turned off when upgrading a competent account, and the administrator is involved in forming an expert editor based on the recommendations provided by the system. Other levels are automatically upgraded based on the points earned through the correct marking of records. Next, the reporter can submit phishing records in PhishRepo, either manually or automatically. The beneficiary user connects with PhishRepo to download phishing data only. Since the data distribution process needs a registered user type, the beneficiary type is added to the proposed solution. Consequently, the guest user can only view the public information related to phishing records and use search facilities to search available phishing web pages in the repository. Fig. 2 shows the landing page of PhishRepo that is visible to all the accounts mentioned above.

A valid email is required when creating a PhishRepo account, and the administrator is responsible for the final confirmation of a new account. Since a human user or an anti-phishing tool could become a reporter, the reporter account has two subscriptions: individual or corporate. The anti-phishing tools always act as a corporate reporter, and those accounts own an application key for authentication. Further, the cor-

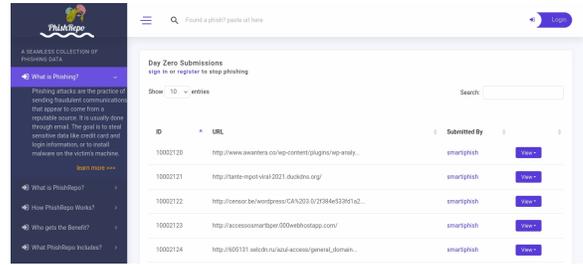


Fig. 2. The Landing Page of PhishRepo

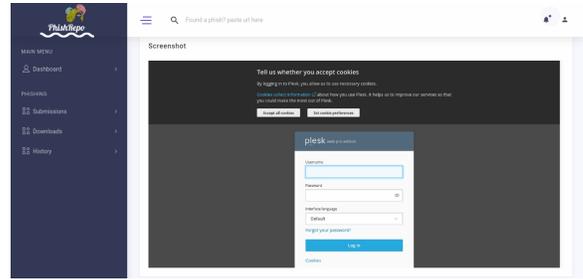


Fig. 3. A Captured Screenshot of a Visible Web Page

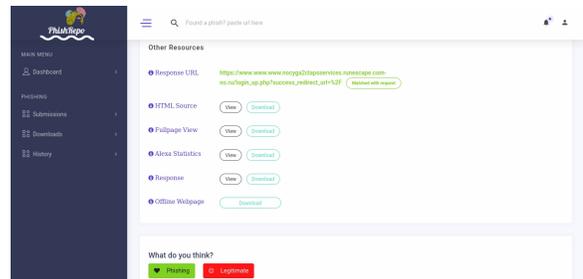


Fig. 4. PhishRepo Displays Additional Information Sources under the Other Resources Section. The Registered user can Download or View that Information Relevant to a Submission.

porate account has an optional field called ‘return address’, which sends daily updates about the submissions. Section IV-C discusses this option in detail.

PhishRepo uses two authentication methods. 1). A login form 2). An application key. The first method is the standard approach in most scenarios, and it uses a username and a password for the verification process. The second one is only applicable to corporate accounts like anti-phishing tools and is only used when an automatic submission is processed with PhishRepo. In that case, the corporate account user must follow the predefined data format to submit a phishing record.

However, one request can submit only one phishing record, and multiple requests are required for numerous submissions. After the authentication is done, it checks for duplication within the repository through URL string matching, and if no duplication is found, the URL is added to the Initial Phishing Records (IPR) queue alongside the submitter information. In PhishRepo, the IPR queue is a double-ended queue (deque) that uses first-in, first-out (FIFO) logic.

2) *Accumulation*: This component is crucial in PhishRepo since it is responsible for the data collection. It starts once a request comes from Server A (Fig. 1). First, it tries to download the complete web page of the submitted URL. If the download process fails due to any exception, the accumulation component skips that URL and moves to the next since the web page is a vital and mandatory information source in PhishRepo. Meanwhile, the data collection process captured the response details and saved them under the record because, in some cases, there can be some mismatches in request and response details which may be helpful in the verification process. After the web page download is done, the screenshot of the web page is captured in full view and visible level (Fig. 3). Then, the possible third-party information is downloaded. This information is kept as additional information (Fig. 4) in PhishRepo and is not mandatory due to the service limitations that exist in the third-party services.

3) *Deduplication*: Deduplication is crucial in PhishRepo that eliminates redundant data such as duplicate phishing pages with the same target. At the beginning of the input module, the authentication component takes the necessary actions to eliminate duplications based on the URL. However, the different URLs do not guarantee that duplications could be avoided in a phishing dataset since most phishing pages are created using phishing kits [4] and released to the public. Therefore, a dataset could have different URLs for similar page structures, as shown in Fig. 5 and make duplicates to machine learning processes that cause data leakage at the end.

In that context, PhishRepo's deduplication component is responsible for eliminating such duplicates with the support of the Perceptual Hashing (pHash) technique [49] that was exercised in the literature for similar scenarios [50]. The component handles the elimination of duplicates as an inline task that affects before saving the accumulated phishing records to the local storage or database, as shown in Fig. 1. Therefore, none of the submissions is identified as duplicate records kept in PhishRepo's repository.

The deduplication process depends on the visual level screenshot downloaded during the accumulation process. It uses the pHash technique to determine the similarity of two phishing pages, and PhishRepo maintains a list of hashes computed for the saved records. During the filtering process, a perceptual hash value is first generated for the newly captured screenshot and compared with the already stored hash values to check whether the new one is a duplicate of an already saved web page. The comparison is made through the distance factor (d) calculated using the two instances' hash values. However, if an exact matching is found ($d = 0$), one of the records will be removed from the repository to address the redundancy factor. In that case, which one to eliminate is dependent on the PhishRepo's setting called 'Dedup Action'. The Dedup Action has two values: new and old. If the value 'new' is enabled, the component will save the new record and remove the old one from the repository, and in the other way, it is not going to save the new record, and the old will remain. The setting is introduced just to have a flexible deduplication process within the PhishRepo, and the administrator is responsible for activating a specific Dedup Action to have a diverse phishing data collection.

Since the deduplication process depends entirely on

the screenshot captured during the accumulation process, PhishRepo is configured to check for near-duplicates for a given period to eliminate any page loading issues during screen capturing. However, it is not practical to check near-duplicates of a new screenshot with all the available records in PhishRepo since the comparison process takes time. Therefore, the deduplication component checks for near-duplicates only for the last three days since most phishing attacks end within three days [51], [31]. Then, theoretically, PhishRepo assumes that the near-duplicates that may exist out of the three days are different phishing attacks.

However, there should be an optimal distance threshold (d_α) for a meaningful selection for the near-duplicates. Therefore, d_α is selected based on 1,000 random samples from an older version of PhishRepo that does not include the deduplication component. Then, pHash values of the screenshots available in the sample were computed, and each pair's d values were calculated. After that, a manual investigation was carried out to examine the accuracy and noted that the accuracy of the similarity of a pair had been decreased drastically when d became more than 10, as shown in Fig. 6. Therefore, d_α was selected as 10, and $0 < d < 10$ are considered near-duplicates in PhishRepo. However, the near-duplicates elimination process does not affect the Dedup Action setting, and if a near duplication is found, the new submission will be entirely discarded from PhishRepo to maintain a diverse phishing data repository.

4) *Targeted Attack Prevention (TAP)*: The main intention of PhishRepo is to collect phishing data to strengthen future anti-phishing tools against phishing attacks. That intention creates opponents (i.e., phishers) to PhishRepo. Therefore, PhishRepo may become a victim of some targeted attacks to disrupt the data collection process of the repository. The denial of service (DoS) attack is a possible threat [36], and there can be other specific attacks like false data injections. However, the network architecture presented in Fig. 1 strengthens the network level protection to a certain extent, and the implemented TAP component provides application-level protection to PhishRepo. The TAP component uses four strategies to have additional application security other than the standard security practices.

- Application key-based authentication – only the users with an application key can submit records automatically
- High-volume restriction – limit number of submissions from one corporate account
- Maximum IPR queue length – limit the number of request processes by the accumulation component
- False ban – Bans the reporters who have falsely recorded submission trend

As described in the authentication component, all the reporters should have an application key when submitting a phishing record automatically to PhishRepo. It limits the attacking trend since the attacker must obtain a valid application key to enter the system. If an attacker comes with a proper application key, then the subsequent countermeasures try to minimise the impact of those attacks. First, a high-volume restriction policy is implemented in PhishRepo to submit only



(a) <https://cbahospitalar.com.br/002WG/well-fargo-RD528-detail/>



(b) <https://mail.cbahospitalar.com.br/002WG/well-fargo-RD528-detail/>

Fig. 5. Different URL Examples for the Same Phishing Target

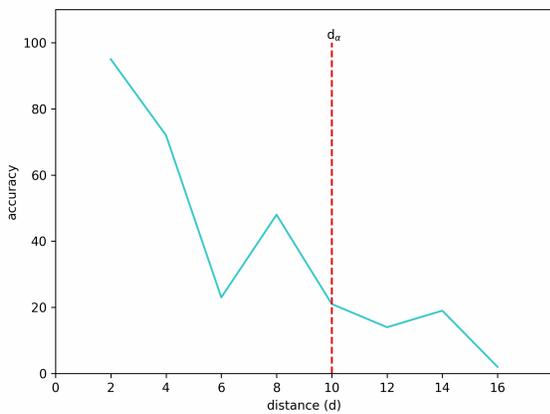


Fig. 6. Estimation of Distance threshold d for Near-Duplicate Detection

a limited number of records for a given period from one reporter. Since the chance of seeing a phishing URL is lower than a legitimate one [22], there is no possibility of submitting many records within a given period since PhishRepo requires real-time submission and discourages batch processing. As previously mentioned, the submissions come as requests, and one request carries only one submission. Therefore, the number of submissions equals the number of requests from a reporter side. Then, if he exceeds the limit (i.e., ten requests per minute), his account is automatically banned for five minutes, and if it is frequent, the account is blocked permanently by the TAP. Further, TAP is responsible for reporting abnormal behaviour to the administrator, and the administrator can take necessary actions on those.

In a specific situation, if an attacker bypassed both the mentioned countermeasures, the maximum IPR queue length is used to maintain a fixed-length queue to avoid overloads of the memory. Then there may be no performance hits, and PhishRepo may function without interruption. However, since the accumulation component takes URLs from the IPR queue, some submitted records may be removed without processing in a special attack. Although it seems wasted, PhishRepo does not intend to collect all the submitted phishing records and work only with the possible submissions when expanding the available phishing records.

In addition, the false ban strategy is another security consideration used in PhishRepo to avoid wrong data injections. The incorrect data may damage the proposed solution's trustworthiness and waste many resources. Therefore, PhishRepo is used to verify whether the collected phishing records are correct. That process is discussed in detail under Section IV-B. This false ban strategy checks the validity of the submitted phishing records week by week for each reporter and calculates an accuracy percentage. If the rate is less than a defined threshold value for an account, that account is suspended automatically and reported to the administrator for further actions.

5) *Manual Submission*: This component is implemented to cater to the generic manual submission needs. However, manual submission is not entertained in PhishRepo since real-time phishing records are required to store correct information sources. In some cases, a manual submission may be a particular need. Therefore, this component is added to the PhishRepo. Manual submission is a simple component, and the primary responsibility is to collect the phishing URL from the interface and send it to the accumulation module to process it further.

B. Verification Module

PhishRepo verifies all the submitted phishing records regardless of the source it gets. It is a two steps process named alpha verification and beta verification. Fig. 7 represents the workflow of the verification process, and the main two steps are explained in detail in the following sub-sections.

1) *Alpha Verification*: It is the first verification done by the PhishRepo after a record is successfully added to the repository. This alpha verification is done using two popular phishing verification solutions in the current context: Phish-Tank and Google Safe Browsing (GSB) [30]. These solutions have free Application Programming Interface (API) support to get information about phishing sites. Therefore, the collected URLs are submitted to both these services. If one or both marked the submissions as phishing, the verification module flags the relevant records as 'verified'. When the verification solutions do not provide any result for a specific submission, in that case, that record is marked with a 'processed' flag. It indicates that the alpha verification is processed on the record

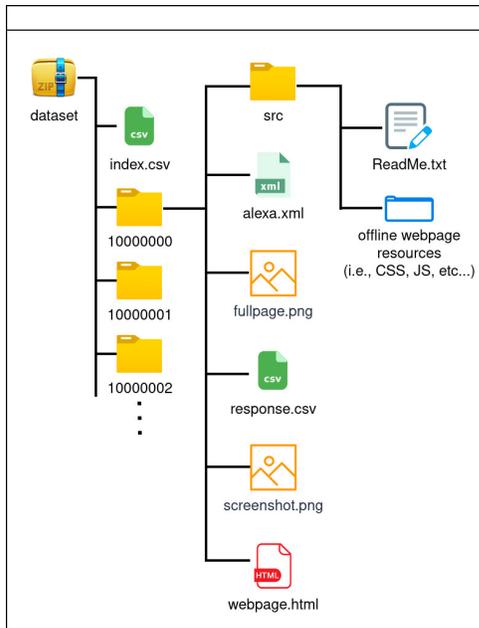


Fig. 10. The Hierarchical Structure of the Zip File

search engine?. The ultimate goal of these questions is to give a second chance to the editor to think about the decision before submitting it to PhishRepo.

C. Distribution Module

Data distribution is the primary goal of PhishRepo. Therefore, the distribution module plays a vital role in the proposed solution. However, the distribution module is only available for registered accounts, and as mentioned before beneficiary user type is specifically designed to support this process. PhishRepo provides several diverse information sources in raw format for each download. Those are the HTML page, visible level and full-page screenshots, response return for the made request, Alexa statistics and offline web page. Further, PhishRepo facilitates the data distribution in two ways: user queries and reporter subscriptions.

1) *User Queries*: The registered users can log in to PhishRepo and query for phishing data. PhishRepo requests data duration and the information sources required by the user. The full dataset could be downloaded from a separate menu item, and it includes all verified phishing records available in PhishRepo. The user query is either for total data or selected data; the final output of the download process is a zip file that includes an index file for easy navigation. Then it is available for users to download. Fig. 10 shows the hierarchical structure of the downloadable zip file.

However, there could be situations like the information sources missed in some folders due to exceptions in the accumulation process. In that case, the index file is vital to find out what is missing since it has columns like an eight-digit number that holds the mapping between the index file and the dataset folders, the request URL, the response URL, the data collection date, and attributes indicating the presence of the visible level screenshot, full-page view, Alexa statistic file, the offline web page, and response header file.

TABLE I. DETAILS OF THE USED PHISHING DATASETS

Dataset Name	Number of Data	
	Phishing	Legitimate
PhishRepo	5,275	0
Ex-PhishRepo	2,029	0
Web2Vec [19]	21,296	24,800
PhishPedia [38]	29,048	22,252

2) *Reporter Subscription*: The reporter subscription module's primary purpose is to give corporate reporters a unique benefit for their vital contributions. So far, it is clear that the corporate reporter is the key user who runs the proposed solution for the long term. Therefore, the PhishRepo is designed to automatically send feedback on what they have reported to the repository to encourage and admire the corporate reporter. As mentioned earlier, the reporter account has a particular field called 'return address'. PhishRepo uses this field value and sends daily feedback to the corporate reporter for their submission. However, feedback for a particular record waits until PhishRepo confirms the records label and keeps track of the sent feedback to avoid any duplication of feedback. The feedback report is sent as a CSV file in PhishRepo.

V. EXPERIMENTS AND RESULTS

The study used two main experiments to evaluate PhishRepo from diversity and its effectiveness in machine learning-based anti-phishing studies. Four primary datasets were used in those experiments, including two recently used public phishing datasets that include the URL, HTML page and screenshot of the relevant phishing instances.

A. Datasets

The four datasets used in the experiments are Web2Vec, PhishPedia, Ex-PhishRepo and PhishRepo. Table I presents the details of those datasets.

The PhishRepo and Ex-PhishRepo datasets were downloaded from the online phishing data repository presented in this paper. However, the Ex-PhishRepo dataset was downloaded before the deduplication component (Section IV-A3) was introduced to the proposed solution. Therefore, duplicate or near-duplicate phishing web pages were not filtered in the Ex-PhishRepo data. The PhishRepo dataset was downloaded after the deduplication filter was influential in the presented work. Therefore, the impact of the filter should be visible in the PhishRepo dataset. The initial level phishing URLs for both the Ex-PhishRepo and PhishRepo datasets were downloaded from PhishTank and OpenPhish. Therefore, the phishing data available in both datasets were valid phishing instances, and both datasets were available online [52] for further reference. Moreover, the Ex-PhishRepo dataset data were collected by the PhishRepo system from 29 September 2021 to 17 October 2021, and the PhishRepo dataset data were collected from 23 October 2021 to 02 February 2022.

The Web2Vec dataset is an online phishing dataset (<https://github.com/Hanjingzhou/Web2vec>) recently used by [19] when developing their anti-phishing solution. The dataset contained 21,303 phishing instances from PhishTank from September 2019 to November 2019 [19]. However, the current

work only could use 21,296 instances out of the total phishing instance of the dataset due to some data extraction issues. Similarly, the PhishPedia dataset is also a recently used phishing dataset by [38]. It contained 29,496 phishing web pages, and OpenPhish's premium account was used when downloading those data. The authors have publicly shared the dataset and are available online (<https://drive.google.com/file/d/12ypEMPRQ43zGRqHGut0Esq2z5en0DH4g/view?usp=sharing>) for anyone to download. The study could only use 29,048 phishing items from the PhishPedia phishing dataset since few data items reported some issues during the extraction.

Since the proposed PhishRepo solution distributes only phishing attack-related data, the PhishRepo and Ex-PhishRepo datasets do not contain legitimate data, as shown in Table I. However, recent anti-phishing studies already used the Web2Vec and PhishPedia datasets. Therefore, both these datasets were attached legitimate data used by those studies, and Alexa was the source for legitimate data in both cases.

B. Diversity of PhishRepo

The main objective of PhishRepo is to provide a diverse phishing dataset for machine learning-based anti-phishing studies. Therefore, PhishRepo output was evaluated from different perspectives to check whether the proposed solution achieved a diverse dataset. However, there is no widely accepted method to check the diversity of a dataset [22], but [22] have proposed two main criteria to use when measuring the diversity of a phishing dataset. Those are the number of different domains and the number of different top-level domains (TLDs). However, literature has shown that the HTTPS based phishing attacks and URL character length distribution are also essential to consider in the current phishing attack nature to have unbiased, accurate model training at the end [53], [7].

Even though these four could be taken as standard criteria to check the diversity, none of the studies in the literature considered the tendency of data leakage in a dataset that has been discussed in Section IV-A3. However, the current study has identified it as an essential factor and used it as the fifth criterion to check the diversity of the PhishRepo dataset. Although Table I presents four datasets, the PhishRepo and Ex-PhishRepo datasets had the exact behaviour in one to four experiments since the deduplication filter was the only noticeable difference in those two datasets. Therefore, the Ex-PhishRepo dataset was not used as a separate dataset during one to four experiments, and it was effectively used in experiment five to show the impact of the deduplication filter.

1) *Distribution of Domains and TLDs*: The domain and TLDs distribution of a dataset depends on the URL of the phishing page. Therefore, the study first extracted unique domains and TLDs from each dataset. Then frequencies of those were calculated separately. After that, the top fifty domains and TLDs were selected from each dataset. Finally, the percentage of the selected domains proportionally to the size of the relevant dataset was calculated, and those values were plotted in ascending order to have the relevant distribution. Fig. 11 and 12 shows the distribution of domains and TLDs, respectively.

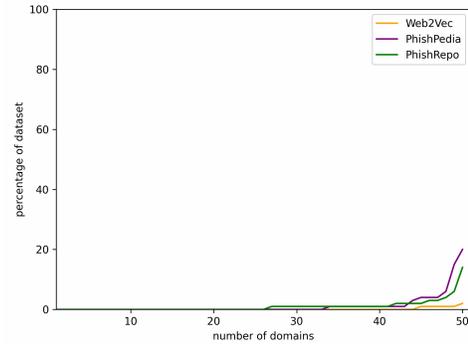


Fig. 11. Distributions of Domains in each Dataset

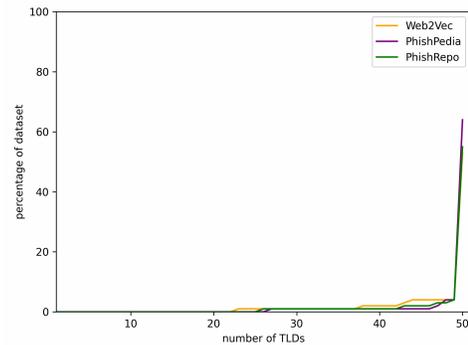


Fig. 12. Distributions of TLDs in each Dataset

As shown in Fig. 11, all the datasets have used more than fifty different domains and TLDs. PhishPedia had a high percentage of the same domain among all datasets, and it was 20% relative to the entire dataset. The PhishRepo dataset had 14% of the same domain, and Web2Vec recorded the lowest number of the same domains. However, the situation is slightly different regarding TLD distribution, as illustrated in Fig. 12. The three datasets have used more than 50% of the '.com' TLD, and it is acceptable because the popular TLDs like '.com' are more often used in phishing nature [22]. Although '.com' acquired a high percentage in all three datasets, more than 50 different TLDs have been included in Web2Vec, PhishPedia, and PhishRepo datasets. Such distribution in domains and TLDs signifies a diverse dataset [22]. Therefore, PhishRepo's dataset is diverse in the perspective of the distribution of domains and TLDs.

2) *URL Character Length Distribution and Percentage of HTTPS*: The current anti-phishing domain is more toward representation learning approaches like deep learning [12], [17], [18], [19], and it results in black box models that do not visualize the features used during the decision making [12]. Therefore, if a phishing dataset does not have a standard distribution in URL character length as presented in [14] work, it may result in inadequate models for real scenarios [53]. Further, as shown in the APWG report [7], more than 80% of present phishing attacks have come with the HTTPS label, indicating that a high percentage of HTTPS in a phishing dataset is also vital to have a realistic scenario during the model training. Therefore, the number of characters in a URL and

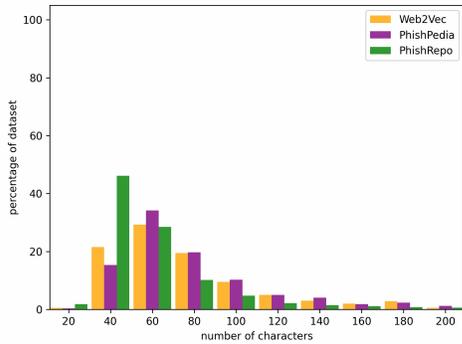


Fig. 13. Character Length Distribution in each Dataset

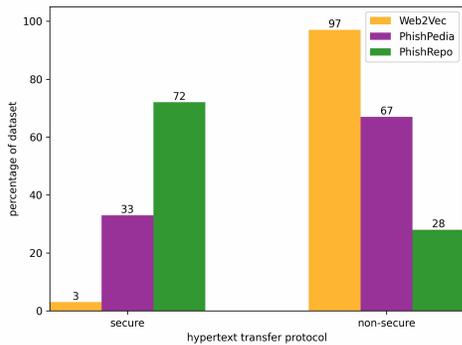


Fig. 14. Percentage of Secure Phishing URLs in each Dataset

percentage of secure phishing URLs proportionally to the size of the relevant dataset was calculated to have URLs character length distribution and percentage of secure phishing URLs.

According to Fig. 13, the URL character length in all three datasets has shown a standard distribution. The highest number of PhishRepo URLs belonged to the 20 to 40 character length category. The other two datasets had a high percentage of 40 to 60 character length URLs. However, all three datasets had URLs under different categories, indicating that these three datasets are diverse in terms of URL character length. In contrast, the secure URLs were deficient in the Web2Vec and PhishPedia datasets. Fig. 14 visualised that the Web2Vec and PhishPedia datasets had 3% and 33% secure URLs. However, current statistics highlighted that nearly 80% of phishing URLs are used HTTPS in the current phishing context [7]. Since it is not reflected in the Web2Vec and PhishPedia datasets, it may lead to inadequate models when these datasets are used in training. However, PhishRepo is shown a high percentage of secure phishing instances, and it has more than 70% of the used dataset. It indicates that the PhishRepo dataset is up to date, and the present phishing nature is sufficiently absorbed.

3) *The Tendency of Data Leakage:* Data leakage is one of the leading machine learning errors and results in poor prediction outcomes. It happens when the information used in the model train appears during testing time. In the context of phishing data, this can happen in two ways. First, the same data is used multiple times, like the phishing website <https://xyz.com> appears on many occasions in the dataset. Next, it can happen due to different URLs for the same

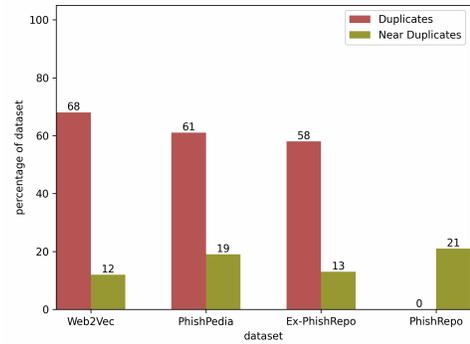


Fig. 15. Percentage of Duplicates and Near-Duplicates of each Dataset

phishing website, as shown in Fig. 5. In both cases, data leakage could happen if the percentage of duplication pairs is high. Therefore, the current study's data leakage tendency is measured based on the number of duplication pairs available in the used datasets. However, none of the previous studies used such a test to check the tendency of data leakage, which may be the first of its kind.

The experiment used the captured screenshots of phishing web pages since the phishers are always trying their best to have a similar fake page compared to the target page. Therefore, the current study assumed that web pages with similar appearances have the same HTML structure. The assumption is valid in most cases since most phishing websites are coming through phishing kits nowadays [4]. However, the failure of the mentioned assumption will not affect the result of the experiment since the tendency of data leakage is calculated using the number of duplicates. Further, the experiment used a commercial tool called pixolution Flow (<https://pixolution.org/>), an AI-powered visual search engine for managing and searching visual data when finding duplicates and near-duplicates. The pixolution Flow has a docker image that could index up to 5,000 images free, and that docker is used during the experiment. Therefore, 5000 random samples were selected from each dataset using the pixolution docker before starting the experiment.

The experiment used a 1.0 threshold when searching for duplicates, and the near-duplicates search was configured to use a 0.9 threshold since it is the recommended threshold by the tool. The duplicates and near-duplicate percentages of each dataset are presented in Fig. 15. However, during the indexing step of the tool, the Web2Vec dataset could not index all the screenshots listed since twenty-two images had some issues. Therefore, the presented percentages of the Web2Vec dataset are calculated from 4,978 data items.

After introducing the deduplication component, PhishRepo has improved by decreasing the duplicate images to 0 and keeping the near-duplicate percentage around 20, as shown in Fig. 15. Further, the experiment shows that the other datasets, Web2Vec, PhishPedia and Ex-PhishRepo, have higher duplication percentages (Fig. 15) than PhishRepo. Therefore, the study can claim that the present version of the PhishRepo dataset does not tend to leak data since it does not contain any duplicate pairs. However, phishing cannot eliminate the near-duplicates since phishers mainly target popular web-

TABLE II. ANTI-PHISHING SOLUTIONS USED IN THE EXPERIMENT

Solution	Description
URLNet ^a [54]	A deep learning approach that detects malicious URLs directly from the URL.
StackModel ^b [14]	Detect phishing attacks with the support of URL and HTML content features.
HybridDLM ^c [16]	A deep learning model uses direct URLs with manually extracted HTML content features.

^a<https://github.com/Antimalweb/URLNet>

^bhttps://drive.google.com/drive/folders/1T4uHRxb_Uk5_kXcJrq68mZ-ezWSQgs_e

^c<https://github.com/sna-hm/HybridDLM>

sites, and those attacks may have slight differences. Although PhishRepo’s deduplication filter is configured to discard near-duplicates, it checks near-duplicates only three days from a given date due to the previously mentioned practical limitations (Section IV-A3). Therefore, this experiment concludes that the PhishRepo dataset is well-suited for machine learning-based anti-phishing tasks from the perspective of data leakage.

Additionally, based on the experiments mentioned above, the study has shown that the proposed solution, PhishRepo produces a diverse dataset, and it is more suitable for machine learning-based phishing detection studies.

C. PhishRepo’s Effectiveness in Anti-Phishing Studies

The ultimate goal of PhishRepo is to provide a phishing dataset for machine learning-based anti-phishing studies. Therefore, a different experiment was performed to prove the effectiveness of PhishRepo’s output compared to recently used public phishing datasets. The Web2Vec and PhishPedia datasets were selected for this purpose since both are similar to a certain extent to the PhishRepo dataset from the perspective of the available information. Further, these two datasets were already exercised with two recent anti-phishing solutions [19], [38] that have shown high performances. Therefore, these two datasets, alongside the PhishRepo dataset, were used to train several existing machine learning-based anti-phishing solutions (Table II) separately and evaluated those against the latest phishing attacks.

1) *Train and Test Datasets:* PhishRepo is an online phishing data repository that expects to grow with time. Although PhishRepo is in the early stage of its journey, it managed to collect around 5000 latest phishing data, and this experiment was planned with these data to compare the effectiveness of PhishRepo data with the state of arts phishing datasets.

Generally, a machine learning model needs a training and test dataset. Therefore, as the first step, the required datasets were constructed. However, the primary intention of the proposed PhishRepo solution is to produce the latest phishing data for anti-phishing studies. Therefore, the experiment required the latest phishing data for the evaluation process. Out of all the selected datasets, the PhishRepo dataset had the latest phishing attacks since it collected phishing attacks up to 02 February 2022. Therefore, the last ten days of phishing attacks were initially separated from the PhishRepo dataset and had 518 records. Those 518 records were added to the test dataset under the phishing label, and the remaining data (i.e. data up to 21 January 2022) were selected as PhishRepo’s training dataset.

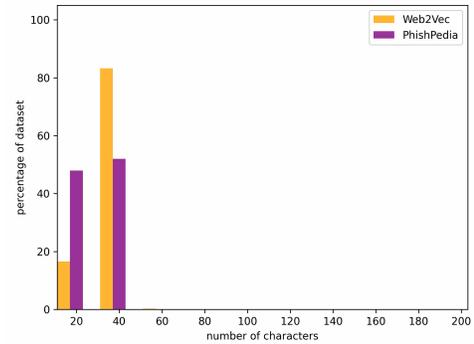


Fig. 16. Legitimate URL Character Length of Web2Vec and PhishPedia

It had 4,757 data, and the amount is reasonable to train a machine learning-based anti-phishing solution since [18] also did a successful anti-phishing study with 4,700 total phishing instances.

The experiment was planned to change only the phishing examples during the training. Therefore, other factors such as the dataset’s size, the legitimate examples seen by the solutions and the test dataset, including phishing and legitimate, were kept constant. Since the number of phishing examples needs to be the same in all three cases, 4,757 phishing data were randomly selected from the Web2Vec and PhishPedia datasets to construct the Web2Vec and PhishPedia training datasets.

Next, the experiment required legitimate examples for effective learning. Table I shows that the Web2Vec and PhishPedia original datasets had a legitimate collection. However, those legitimate data were collected from Alexa. If a legitimate dataset is constructed using Alexa without a specific strategy and mixed with a phishing dataset collected from PhishTank, the URL character length plays a significant role and may produce malfunctioned classifiers [53], [22]. Therefore, the Web2Vec and PhishPedia datasets were initially examined by plotting the character length of available legitimate URLs. Fig. 16 shows the character length distribution of those legitimate URLs. It visualises that the mentioned URL character length issue exists with both Web2Vec and PhishPedia legitimate data compared with phishing URL character length available in Fig. 13. Since it affects the final evaluation process of the planned experiment, Web2Vec and PhishPedia legitimate data were not used to construct the training dataset. Therefore, an online phishing dataset named Phishing Websites dataset [55] was used to collect the required legitimate data since it had a reasonable URL character length distribution compared to the [14] work, as shown in Fig. 17.

The experiment planned to have a balanced dataset during the training. Therefore, 4,757 and 518 legitimate data were randomly selected from the Phishing Websites dataset for the training and test datasets. Finally, the training and test datasets contained 9,514 and 1,036 data. Since the experiment wanted consistent legitimate data to effectively evaluate the PhishRepo dataset performance, the same legitimate training samples were added to the Web2Vec, PhishPedia and PhishRepo training datasets. The test dataset was similar in all the experiments, and it has used to evaluate the selected model’s performance in each case.

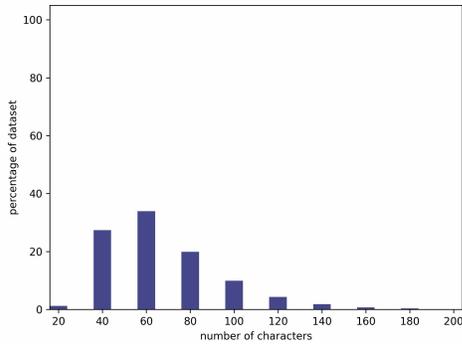


Fig. 17. Legitimate URL Character Length of Phishing Websites Datasets

TABLE III. TRAINED MODELS' PERFORMANCES WITH EACH DATASET

Solution	Dataset	Accuracy	f1-Score	FNR*
URLNet	Web2Vec	72.20%	70.79%	0.326
	PhishPedia	78.09%	77.72%	0.236
	PhishRepo	82.24%	83.45%	0.104
StackModel	Web2Vec	58.11%	36.55%	0.759
	PhishPedia	75.87%	69.96%	0.438
	PhishRepo	89.00%	88.97%	0.112
HybridDLM	Web2Vec	67.18%	51.98%	0.645
	PhishPedia	88.22%	87.07%	0.207
	PhishRepo	93.92%	93.89%	0.066

*False Negative Rate (FNR): The number of phishing instances is marked as legitimate proportionally to all existing phishing instances.

2) *Performance Evaluation:* First of all, Table II anti-phishing solutions were separately trained using the Web2Vec, PhishPedia and PhishRepo training datasets. Then, these models were evaluated using the test dataset. The obtained results during the evaluation process are shown in Table III. However, the URLNet experiment had different models based on the used embedding values, and the best-performed model was selected to present the final results.

As shown in Table III, the PhishRepo dataset has shown high accuracies and f1-score and low FNR in all three cases where the PhishRepo dataset was used. Since the datasets size, legitimate examples, and test dataset were constant in all cases, the phishing examples made the performance difference in each dataset. It implies that the models effectively learned most phishing scenarios with the PhishRepo dataset, and it is more effective when presenting phishing examples for the used models.

Although the PhishRepo dataset has shown some significant performance, one might argue that it is due to the latest phishing examples it contained. However, that is the main objective of the current work. The machine learning-based anti-phishing studies lack the latest phishing data. Therefore, most models are trained with old phishing data, and those models may not perform well with the latest phishing attacks since phishing characteristics constantly change over time. That is what exactly happened during the performed experiment. Since the test dataset contained the latest phishing attacks and both Web2Vec and PhishPedia had old phishing examples, current significant phishing characteristics might not be captured during the training. However, PhishRepo is constantly collecting these phishing examples. Therefore, it contained the latest

phishing examples, and more significant characteristics are existed in PhishRepo examples to detect the latest phishing attacks. Thus, the model trained with the PhishRepo dataset captured these new characteristics and performed well with the latest attacks.

Furthermore, as shown in Fig. 15, the PhishRepo dataset does not contain duplicate phishing examples. However, in Web2Vec and PhishPedia, duplication is over 50%. Therefore, compared to the PhishRepo dataset, the Web2Vec and PhishPedia datasets may contain fewer unique phishing examples. Then, although the training dataset size is equal, the amount of learning a model can gain through the training set becomes lower in the other two datasets than in the PhishRepo dataset due to the high duplicate phishing instances available. Therefore, PhishRepo output is more suitable for machine learning-based anti-phishing studies since it produces the latest diverse phishing examples for an effective learning process.

VI. DISCUSSION

Machine learning-based phishing detection desires labelled phishing data at present. The unavailability of such data directs anti-phishing research into many challenges. Some of them are, lingered data collection, data obsolescence, low-quantity data, low-quality data, and lack of multi-modal feature representation. These challenges result in inept learning models, weakening the effort to combat phishing. Therefore, it is essential to fill the current gap in the anti-phishing domain to strengthen future detections. As a result, PhishRepo is introduced as a gap-filling solution to deliver future phishing data needs in the anti-phishing domain. However, it is not just a way of storing data; it is responsible for the latest quality data dissemination to enrich the effectiveness of the anti-phishing solutions.

Phisherman [36], [35] is the only solution in the literature with the same aim as PhishRepo. However, PhishRepo is conceptually superior to Phisherman in many design aspects. Some examples include a deduplication filter, a crowdsourcing-based verification process, malicious submission detection, and the ability to report objections. PhishRepo generally benefits from automated submission architecture, and its design allows it to access a variety of information sources in raw format. Furthermore, the deduplication filter ensures diverse data collection and the elegant verification process used in PhishRepo results in high-quality data. The objection reporting helps to maintain the quality even more. Moreover, the innovative data distribution structure is purposely designed in PhishRepo to attract users, primarily autonomous anti-phishing tools. These tools could integrate PhishRepo more thoughtfully to handle the constantly changing phishing attacks. Further, the proposed network architecture and TAP strategies are critical for the solution's smooth operation from a security perspective.

PhishRepo is a phishing data repository that is accessible online. As a result, anyone interested in the solution could gain access to it and obtain the final benefit, the data. The primary audience for PhishRepo is anti-phishing researchers. They can use this solution effectively to eliminate the phishing data hassle. Since the repository includes multi-modal features, the researcher could use PhishRepo to take their research in a new direction. Furthermore, the raw format in PhishRepo supports

representation learning approaches such as deep learning to design differentiated anti-phishing solutions. It is further intended to support reinforcement learning (RL) environments because PhishRepo includes an interactive feedback facility. With this facility, an implemented RL environment could submit its actions for specific observations and receive quality feedback from PhishRepo. Therefore, the data collected by PhishRepo could aid anti-phishing researchers in various ways, allowing them to conduct more effective research. In addition to such, PhishRepo is an excellent solution for data drifting, which mainly affects machine learning models' performance [21], [22]. Therefore, the latest data collected by PhishRepo could be used to retrain existing models to retain their performance in the fast-evolving nature of phishing.

Moreover, PhishRepo is the first study to examine the tendency of data leakage in a phishing dataset in the anti-phishing domain. It found that the deduplication filter introduced in this study causes no data leakage. The experiments conducted to demonstrate the efficacy of the PhishRepo data have also demonstrated that the data are diverse and do not contain duplicate data, which could lead to a data leakage problem. Furthermore, PhishRepo has been compared with two recently used public datasets using three anti-phishing solutions. There also, PhishRepo outperformed other datasets by achieving high detection accuracy, f1-score and low FNR by showing the strength of the proposed solution.

However, the reliability of PhishRepo is primarily determined by the submissions it receives. Therefore, reporters are essential to the proposed architecture, and corporate reporters are critical because PhishRepo encourages real-time submissions rather than manual ones. Another essential role in PhishRepo is the editor, particularly the crowd user, who is always critical to the success of the beta verification process. However, alpha eliminates the need for a beta. Therefore, PhishRepo assumes that few editors can manage beta verification in the early stages. However, the contributions of the reporters and editors are critical for PhishRepo to continue its process and achieve its ultimate goal.

As a general limitation of the solution, the third-party services' availability is critical in PhishRepo, and the failure of some may affect the solution's continuity. Therefore, PhishRepo expects a collaborative effort against phishing rather than individual combat. Further, a few more anti-phishing communities could be integrated into the alpha verification process to strengthen the alpha process and reduce the human workload in the verification. Moreover, archiving some erroneous pages (e.g., 403 pages, 404 pages, and content not found pages) impacts the PhishRepo data quality. Therefore, additional work will be required in the future to automatically detect erroneous or unwanted pages via web page screenshots and remove such data points from the repository. Then, the quality of the PhishRepo data could be improved further, providing researchers with significantly less noisy data.

VII. CONCLUSION

While machine learning methods are gaining popularity in phishing detection, the lack of labelled data limited the viability of machine learning-based anti-phishing solutions. Large-scale, diverse data sources are essential in phishing

detection in today's context, and it helps researchers have effective machine learning models to combat phishing in the future. PhishRepo comes under these circumstances, and it is an online phishing data repository that collects, verifies, disseminates, and archives real-time phishing data. PhishRepo uses a tactical approach from collection to dissemination. Therefore, it always guarantees the quality of data it saves. Further, automated submission, deduplication filtering, automated verification, crowdsourcing-based human interaction, objection reporting window, and security considerations outperform PhishRepo over similar solutions in the phishing domain.

However, the proposed gap-filling solution's reliability depends on its submissions. Although it is a limitation, PhishRepo identifies its importance and promotes specific tactics to bind users to the solution. Therefore, PhishRepo will be an essential service to provide quality labelled multi-modal feature-based phishing data to detect phishing attacks effectively in the future.

ACKNOWLEDGMENT

The authors acknowledge the support received from the Center for Information Technology Services (CITeS) of the University of Moratuwa, Sri Lanka and Dr Chamath Keppitiyagama of the University of Colombo School of Computing, Sri Lanka.

REFERENCES

- [1] ENISA, *ENISA threat landscape report 2018: 15 top cyber threats and trends*. Publications Office, 2019. [Online]. Available: <https://data.europa.eu/doi/10.2824/622757>
- [2] Z. Alkhalil, C. Hewage, L. Nawaf, and I. Khan, "Phishing attacks: A recent comprehensive study and a new anatomy," *Frontiers in Computer Science*, vol. 3, Mar. 2021. [Online]. Available: <https://doi.org/10.3389/fcomp.2021.563060>
- [3] W. D. Yu, S. Nargundkar, and N. Tiruthani, "A phishing vulnerability analysis of web based systems," in *2008 IEEE Symposium on Computers and Communications*. IEEE, Jul. 2008. [Online]. Available: <https://doi.org/10.1109/iscc.2008.4625681>
- [4] K. L. Chiew, K. S. C. Yong, and C. L. Tan, "A survey of phishing attacks: Their types, vectors and technical approaches," *Expert Systems with Applications*, vol. 106, pp. 1–20, Sep. 2018. [Online]. Available: <https://doi.org/10.1016/j.eswa.2018.03.050>
- [5] H. Huang, S. Zhong, and J. Tan, "Browser-side countermeasures for deceptive phishing attack," in *2009 Fifth International Conference on Information Assurance and Security*. IEEE, 2009. [Online]. Available: <https://doi.org/10.1109/ias.2009.12>
- [6] Z. Dou, I. Khalil, A. Khreishah, A. Al-Fuqaha, and M. Guizani, "Systematization of knowledge (SoK): A systematic review of software-based web phishing detection," *IEEE Communications Surveys & Tutorials*, vol. 19, no. 4, pp. 2797–2819, 2017. [Online]. Available: <https://doi.org/10.1109/comst.2017.2752087>
- [7] APWG, "Phishing activity trends report: 4th quarter 2020," *Anti-Phishing Working Group. Retrieved February, 09*, p. 13, 2021.
- [8] N. C. R. L. Y. Teraguchi and J. C. Mitchell, "Client-side defense against web-based identity theft," *Computer Science Department, Stanford University*. Available: <http://crypto.stanford.edu/SpoofGuard/webspoof.pdf>, 2004.
- [9] S. Sheng, B. Magnien, P. Kumaraguru, A. Acquisti, L. F. Cranor, J. Hong, and E. Nunge, "Anti-phishing phil," in *Proceedings of the 3rd symposium on Usable privacy and security - SOUPS '07*. ACM Press, 2007. [Online]. Available: <https://doi.org/10.1145/1280680.1280692>
- [10] P. Prakash, M. Kumar, R. R. Kompella, and M. Gupta, "PhishNet: Predictive blacklisting to detect phishing attacks," in *2010 Proceedings IEEE INFOCOM*. IEEE, Mar. 2010. [Online]. Available: <https://doi.org/10.1109/infcom.2010.5462216>

- [11] M. Baslyman and S. Chiasson, "'smells phishy?': An educational game about online phishing scams," in *2016 APWG Symposium on Electronic Crime Research (eCrime)*. IEEE, Jun. 2016. [Online]. Available: <https://doi.org/10.1109/ecrime.2016.7487946>
- [12] A. C. Bahnsen, E. C. Bohorquez, S. Villegas, J. Vargas, and F. A. Gonzalez, "Classifying phishing URLs using recurrent neural networks," in *2017 APWG Symposium on Electronic Crime Research (eCrime)*. IEEE, Apr. 2017. [Online]. Available: <https://doi.org/10.1109/ecrime.2017.7945048>
- [13] A. K. Jain and B. B. Gupta, "Phishing detection: Analysis of visual similarity based approaches," *Security and Communication Networks*, vol. 2017, pp. 1–20, 2017. [Online]. Available: <https://doi.org/10.1155/2017/5421046>
- [14] Y. Li, Z. Yang, X. Chen, H. Yuan, and W. Liu, "A stacking model using URL and HTML features for phishing webpage detection," *Future Generation Computer Systems*, vol. 94, pp. 27–39, May 2019. [Online]. Available: <https://doi.org/10.1016/j.future.2018.11.004>
- [15] W. Wang, F. Zhang, X. Luo, and S. Zhang, "PDCRCNN: Precise phishing detection with recurrent convolutional neural networks," *Security and Communication Networks*, vol. 2019, pp. 1–15, Oct. 2019. [Online]. Available: <https://doi.org/10.1155/2019/2595794>
- [16] S. Ariyadasa, S. Fernando, and S. Fernando, "Detecting phishing attacks using a combined model of LSTM and CNN," *International Journal of ADVANCED AND APPLIED SCIENCES*, vol. 7, no. 7, pp. 56–67, Jul. 2020. [Online]. Available: <https://doi.org/10.21833/ijaas.2020.07.007>
- [17] C. Opara, B. Wei, and Y. Chen, "HTMLPhish: Enabling phishing web page detection by applying deep learning techniques on HTML analysis," in *2020 International Joint Conference on Neural Networks (IJCNN)*. IEEE, Jul. 2020. [Online]. Available: <https://doi.org/10.1109/ijcnn48605.2020.9207707>
- [18] C. Opara, Y. Chen, and B. wei, "Look before you leap: Detecting phishing web pages by exploiting raw url and html characteristics," 2020.
- [19] J. Feng, L. Zou, O. Ye, and J. Han, "Web2vec: Phishing webpage detection method based on multidimensional features driven by deep learning," vol. 8, pp. 221 214–221 224, 2020. [Online]. Available: <https://doi.org/10.1109/access.2020.3043188>
- [20] M. Khonji, Y. Iraqi, and A. Jones, "Phishing detection: A literature survey," *IEEE Communications Surveys & Tutorials*, vol. 15, no. 4, pp. 2091–2121, 2013. [Online]. Available: <https://doi.org/10.1109/surv.2013.032213.00009>
- [21] D. Sahoo, C. Liu, and S. C. H. Hoi, "Malicious url detection using machine learning: A survey," 2019.
- [22] A. E. Aassal, S. Baki, A. Das, and R. M. Verma, "An in-depth benchmarking and evaluation of phishing detection research for security needs," *IEEE Access*, vol. 8, pp. 22 170–22 192, 2020. [Online]. Available: <https://doi.org/10.1109/access.2020.2969780>
- [23] A. Butnaru, A. Mylonas, and N. Pitropakis, "Towards lightweight URL-based phishing detection," *Future Internet*, vol. 13, no. 6, p. 154, Jun. 2021. [Online]. Available: <https://doi.org/10.3390/fi13060154>
- [24] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, May 2015. [Online]. Available: <https://doi.org/10.1038/nature14539>
- [25] O. K. Sahingoz, E. Buber, O. Demir, and B. Diri, "Machine learning based phishing detection from URLs," *Expert Systems with Applications*, vol. 117, pp. 345–357, Mar. 2019. [Online]. Available: <https://doi.org/10.1016/j.eswa.2018.09.029>
- [26] L. A. T. Nguyen, B. L. To, H. K. Nguyen, and M. H. Nguyen, "An efficient approach for phishing detection using single-layer neural network," in *2014 International Conference on Advanced Technologies for Communications (ATC 2014)*. IEEE, Oct. 2014. [Online]. Available: <https://doi.org/10.1109/atc.2014.7043427>
- [27] E.-S. M. El-Alfy, "Detection of phishing websites based on probabilistic neural networks and k-medoids clustering," *The Computer Journal*, vol. 60, no. 12, pp. 1745–1759, Apr. 2017. [Online]. Available: <https://doi.org/10.1093/comjnl/bxx035>
- [28] W. Chen, W. Zhang, and Y. Su, "Phishing detection research based on LSTM recurrent neural network," in *Communications in Computer and Information Science*. Springer Singapore, 2018, pp. 638–645. [Online]. Available: https://doi.org/10.1007/978-981-13-2203-7_52
- [29] M. Chatterjee and A.-S. Namin, "Detecting phishing websites through deep reinforcement learning," in *2019 IEEE 43rd Annual Computer Software and Applications Conference (COMPSAC)*. IEEE, Jul. 2019. [Online]. Available: <https://doi.org/10.1109/compsac.2019.10211>
- [30] S. Bell and P. Komisarczuk, "An analysis of phishing blacklists: Google safe browsing, OpenPhish, and PhishTank," in *Proceedings of the Australasian Computer Science Week Multiconference*. ACM, Jan. 2020. [Online]. Available: <https://doi.org/10.1145/3373017.3373020>
- [31] V. Zeng, S. Baki, A. E. Aassal, R. Verma, L. F. T. D. Moraes, and A. Das, "Diverse datasets and a customizable benchmarking framework for phishing," in *Proceedings of the Sixth International Workshop on Security and Privacy Analytics*. ACM, Mar. 2020. [Online]. Available: <https://doi.org/10.1145/3375708.3380313>
- [32] A. K. Jain and B. B. Gupta, "A machine learning based approach for phishing detection using hyperlinks information," *Journal of Ambient Intelligence and Humanized Computing*, vol. 10, no. 5, pp. 2015–2028, Apr. 2018. [Online]. Available: <https://doi.org/10.1007/s12652-018-0798-z>
- [33] A. Orunsolu, A. Sodiya, and A. Akinwale, "A predictive model for phishing detection," *Journal of King Saud University - Computer and Information Sciences*, Dec. 2019. [Online]. Available: <https://doi.org/10.1016/j.jksuci.2019.12.005>
- [34] P. Yang, G. Zhao, and P. Zeng, "Phishing website detection based on multidimensional features driven by deep learning," *IEEE Access*, vol. 7, pp. 15 196–15 209, 2019. [Online]. Available: <https://doi.org/10.1109/access.2019.2892066>
- [35] G. Tally, "PhisherMan: A phishing data repository," in *2009 Cybersecurity Applications & Technology Conference for Homeland Security*. IEEE, Mar. 2009. [Online]. Available: <https://doi.org/10.1109/catch.2009.24>
- [36] G. Tally, D. Sames, T. Chen, C. Colleran, D. Jevans, K. Omiliak, and R. Rasmussen, "The phisherMan project: Creating a comprehensive data collection to combat phishing attacks," *Journal of Digital Forensic Practice*, vol. 1, no. 2, pp. 115–129, Jul. 2006. [Online]. Available: <https://doi.org/10.1080/15567280601015564>
- [37] R. M. Mohammad, F. Thabtah, and L. McCluskey, "An assessment of features related to phishing websites using an automated technique," in *2012 International Conference for Internet Technology and Secured Transactions*. IEEE, 2012, pp. 492–497.
- [38] Y. Lin, R. Liu, D. M. Divakaran, J. Y. Ng, Q. Z. Chan, Y. Lu, Y. Si, F. Zhang, and J. S. Dong, "Phishpedia: A hybrid deep learning based approach to visually identify phishing webpages," in *30th USENIX Security Symposium (USENIX Security 21)*. USENIX Association, Aug. 2021, pp. 3793–3810. [Online]. Available: <https://www.usenix.org/conference/usenixsecurity21/presentation/lin>
- [39] K. Beck and J. Zhan, "Phishing using a modified bayesian technique," in *2010 IEEE Second International Conference on Social Computing*. IEEE, Aug. 2010. [Online]. Available: <https://doi.org/10.1109/socialcom.2010.100>
- [40] E. Buber, O. Demir, and O. K. Sahingoz, "Feature selections for the machine learning based detection of phishing websites," in *2017 International Artificial Intelligence and Data Processing Symposium (IDAP)*. IEEE, Sep. 2017. [Online]. Available: <https://doi.org/10.1109/idap.2017.8090317>
- [41] J. C. Chang, S. Amershi, and E. Kamar, "Revolt: Collaborative crowdsourcing for labeling machine learning datasets," in *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. ACM, May 2017. [Online]. Available: <https://doi.org/10.1145/3025453.3026044>
- [42] L. Zhao, G. Sukthankar, and R. Sukthankar, "Incremental relabeling for active learning with noisy crowdsourced annotations," in *2011 IEEE Third Int'l Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third Int'l Conference on Social Computing*. IEEE, Oct. 2011. [Online]. Available: <https://doi.org/10.1109/passat/socialcom.2011.193>
- [43] A. Drutsa, V. Farafonova, V. Fedorova, O. Megorskaya, E. Zermimova, and O. Zhilinskaya, "Practice of efficient data collection via crowdsourcing at large-scale," 2019.
- [44] T. Aitamurto, A. Leiponen, and R. Tee, "The promise of idea crowdsourcing—benefits, contexts, limitations," *Nokia Ideasproject White Paper*, vol. 1, pp. 1–30, 2011.

- [45] A. P. Dawid and A. M. Skene, "Maximum likelihood estimation of observer error-rates using the EM algorithm," *Applied Statistics*, vol. 28, no. 1, p. 20, 1979. [Online]. Available: <https://doi.org/10.2307/2346806>
- [46] P. G. Ipeirotis, F. Provost, and J. Wang, "Quality management on amazon mechanical turk," in *Proceedings of the ACM SIGKDD Workshop on Human Computation - HCOMP '10*. ACM Press, 2010. [Online]. Available: <https://doi.org/10.1145/1837885.1837906>
- [47] D. L. Hansen, P. J. Schone, D. Corey, M. Reid, and J. Gehring, "Quality control mechanisms for crowdsourcing," in *Proceedings of the 2013 conference on Computer supported cooperative work - CSCW '13*. ACM Press, 2013. [Online]. Available: <https://doi.org/10.1145/2441776.2441848>
- [48] C. Eickhoff, "Cognitive biases in crowdsourcing," in *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*. ACM, Feb. 2018. [Online]. Available: <https://doi.org/10.1145/3159652.3159654>
- [49] C. Zauner, "Implementation and benchmarking of perceptual image hash functions," 2010.
- [50] D. T. Nguyen, F. Alam, F. Offi, and M. Imran, "Automatic image filtering on social networks using deep learning and perceptual hashing during crises," 2017.
- [51] R. Gowtham and I. Krishnamurthi, "A comprehensive and efficacious architecture for detecting phishing webpages," *Computers & Security*, vol. 40, pp. 23–37, 2014.
- [52] S. Ariyadasa, S. Fernando, and S. Fernando, "phishrepo-dataset," 2022. [Online]. Available: <https://data.mendeley.com/datasets/ttmmtsgbs8/3>
- [53] R. M. Verma, V. Zeng, and H. Faridi, "Data quality for security challenges," in *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*. ACM, Nov. 2019. [Online]. Available: <https://doi.org/10.1145/3319535.3363267>
- [54] H. Le, Q. Pham, D. Sahoo, and S. C. H. Hoi, "Urlnet: Learning a URL representation with deep learning for malicious URL detection," *CoRR*, vol. abs/1802.03162, 2018. [Online]. Available: <http://arxiv.org/abs/1802.03162>
- [55] S. Ariyadasa, S. Fernando, and S. Fernando, "Phishing websites dataset," 2021. [Online]. Available: <https://data.mendeley.com/datasets/n96ncsr5g4/1>

A Novel Code Completion Strategy

Hayatou Oumarou

University of Maroua, Maroua, Cameroun
LaRI Team, Maroua, Cameroun

Ousmanou Dahirou

University of Maroua, Maroua, Cameroun

Abstract—Programmers rely on a multitude of techniques to speed up the development process. Among these techniques is code completion, a productivity improvement technique widely used by developers to explore APIs and automatically complete a word being typed by providing a progressively refined list of candidate words (or recommendations). Still called auto-completion, it reduces incorrect calls to APIs. Several techniques have been developed to obtain the list of candidates. Some methods use the history of the code, others neural networks or artificial intelligence; some exploit the program's structure through AST. Often the recommendation list is long, and finding suitable candidates comes at a cost. In this work, we propose a strategy that improves the accuracy of recommendation list offered by code completion. We present a sorting approach based on the popularity and importance of the elements (suggestions) of the list by analyzing the usage data of classes, methods, and variables of projects in the same development environment. We implemented our sorting strategy in Pharo (IDE and language), an immersive modern programming environment to show its applicability. The empirical evaluation results of this strategy show that our approach improves the quality of the suggestions.

Keywords—Integrated development environment; code completion; API; code completion tool; pharo

I. INTRODUCTION

Integrated Development Environments (IDEs) have become a critical paradigm for software engineers to speed up the coding process and reduce typos and other common errors. An IDE brings together tools for developing software such as mobile applications, computer or game console applications, web applications, etc. There are more IDEs than programming languages. However, most IDEs are specific to a given language [1]. A modern IDE has various tools distributed together, which are among others: the text editor, the graphical interface, the debugger, the compiler, testing and versioning tools, ... helping the programmer to write code efficiently and accurately by providing it with a set of valuable services such as automatic indentation of code blocks, highlighting of language keywords by color or bold characters, code completion, etc. Code completion is the mechanism allowing from part of a word entered by the user to offer him a progressively refined list of candidates (complements) which could suit the remaining string of characters of the word. This functionality can be found in several applications : text editors (for entering source code, for word processing, etc.), web browsers, as well as specific intuitive input systems installed on smartphones. This work will focus on code completion in IDE editors, highlight its shortcomings, and propose an improvement.

A. Context

Murphy [3] published an empirical study on how 41 Java developers used the Eclipse IDE. One of their findings was that every developer in the study used the code completion feature. Among the top commands executed by the 41 developers, code completion came in sixth with 6.7% of the number of commands executed, sharing the top spots with basic editing commands such as copy, paste, save and delete. Not surprisingly, this has been little discussed: Code completion has become second praxis to implementation activity. Nowadays, every major IDE offers a language-specific code completion system; according to [4] any text editor must provide at least word completion to be considered usable for programming. In the same vein, we did an online survey in June 2020 with local developers. The survey focused on quality practices and measures in software development companies [5].

B. Motivations

Now-a-days, software development has become highly complex and very difficult to master due to the increase in the scale of the projects, the increasingly short development time, the requirements and the quality of the software product. To successfully manage the development of software, it is necessary to consider many parameters, including material resource constraints, the programming languages used, and the human factor. The latter greatly influences the progress of software development. The software production cost includes the hardware cost, the training cost, and the effort required for the development and maintenance. However, the most significant cost is that of the construction effort (software development) because it represents more than 80% of the software production time [6]. Cost overruns and delayed product delivery deadlines are often encountered during software development. To provide an element of the solution to these problems, engineers should strive to act to ease the development process by reducing costs (effort) and development time. Thus, we could increase the productivity of programmers by automating part of the activities (reducing keyboard input by code completion) and by simplifying operations (rapid debugging, less or easily consults the documentation) to achieve the qualitative objectives (quality, cost and time). The code source is the essential element of the software. It is done during implementation and is reviewed during maintenance, representing more than 80% of the software cost. The code completion mechanism takes center stage. Jin [1] highlighted the hidden cost of auto-completion, which mainly impacts developers when code completion techniques produce long recommendations. They show how the length of the recommendations list affects other factors that can lead to inefficiencies in the process. The idea of improving and adapting the completion mechanism is relevant

to provide support to developers by providing considerable time savings and reducing the number of typing errors during development. So software developers will write code more effectively and efficiently.

C. Description of the Problem

Automatic code completion is considered the most used feature in integrated development environments [14]. The recommendations (suggestions) are presented to the programmer during the completion process in a pop-up window in a specific order. A study on the length of self-completion suggestion lists [1] found that around 17% of the lists were 250 items long. In the same study, they showed that the median position of the selected item is beyond the 100th spot. So given the multitude of suggestion possibilities offered by code completion and the way these suggestions are sorted, finding a candidate's place in the list can be tedious or slower than typing in the full name of the element to be completed. Generally, it takes too long to locate the word in the suggestions list. In that case, code completion loses all its importance because, instead of being a tool that increases the coder's productivity by reducing the entry time, it slows down and slow motion. This highlights weaknesses in the suggestion sorting strategies implemented in most code completion systems in development environments. These strategies do not rank the results relevance according to the programmer's context. This work will propose a sorting strategy that improves the precision of the list of suggestions.

D. Structure of the Document

Our main contributions are summarized as follows:

- We propose a new prioritization method.
- We propose a discriminator model on top of the IDE code completion engine that uses contextual scope information for precise code completion.
- We do an extensive experiments showing performance in terms of precision.

The rest of the paper is organized as follows: Section II presents the details of the RBSS strategy. Section III reports the experimental results. Section IV investigates related work. Section V discuss threats to validity of our approach. Section VI concludes the document.

II. THE STRATEGY: RELEVANCE BASED-SORTING STRATEGY

In this section, we present the candidate list refinement strategy called Relevance Based- Sorting Strategy (RBSS). Fig. 1 show an overview of the approach. The idea is to measure the relevance score of each of the list elements (method name, variable name, class name, etc.) according to the number and weight of the links that it receives to refine its position in the suggestions list depending on the context. We need the static and structural information from the source code to do this. Thanks to of the Abstract Syntax Tree (AST) analysis, which returns the syntactic information of the element to be completed, we are able to sort the completion options (names of methods, variables, classes, definitions, etc.) according to their popularity ratings which are the essential

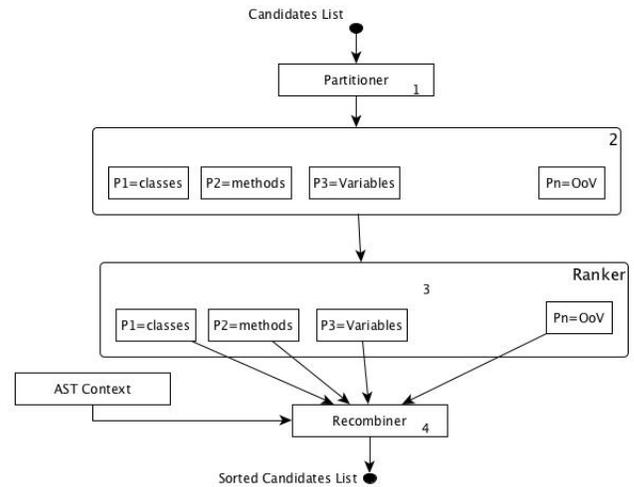


Fig. 1. Overview of RBSS Approach

criteria in referencement. To do this, we contextualize the idea of Google's web page indexing algorithm (PageRank). PageRank is an referencement technique used by the Google search engine to index web pages and provide results that are relevant. The latter, an index used by Google to know the popularity of its index, is noted between 0 and 10. A page is considered very popular if it has a maximum rating or index.

- 1) Partitioning of candidates The first step is to partition the candidates according to the types of elements with the RBSS Partition algorithm as in Fig. 2.

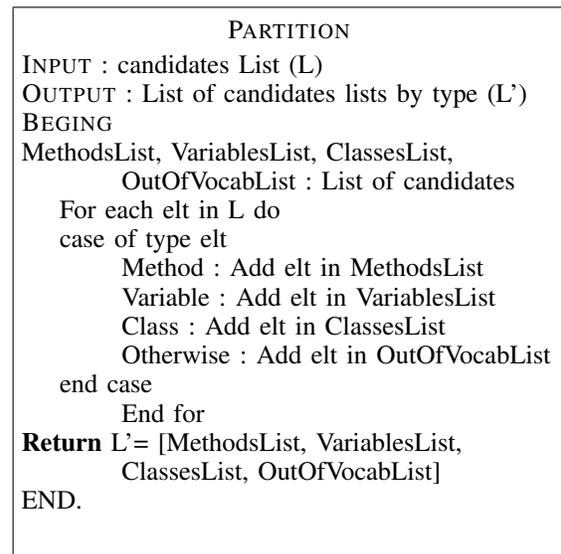


Fig. 2. Algorithm RBSSPartitionner: Partitions a List According to Elements Type

In this step, each item in the list is classified according to:

- Methods For each method in suggestions list, it will be a question of looking at the number of their Senders and their Implementors. The

Senders represent all methods that may use or invoke a given method. The Implementors also work the same way. Instead of returning a list of Senders of a message (or method-envoyeuses), they resolve all classes that implement a method with the same selector. RBSS considers a method popular when many links point to it. Here, a link can represent either a call of the method in question by another method or an implementation of the method in the different classes of the system.

- classes Concerning the case of a class, we will be interested in its references as an attribute for evaluating its popularity. These references are : AllRefInside and AllRefOutside. All-RefInside represents all the references of the other classes of the system which reuse (inheritance, re-implementation,...) the properties of a given class. AllRefOutside does the opposite of AllRefOutside by displaying all the elements of the classes whose attributes a given class uses. Our strategy considers that a class is popular if it has more AllRefInside than AllRefOutside if the size of the AllRefInside list is greater than the size of AllRefOutside.
- variables Regarding the variables, we will consider either the number of methods that use it (outgoingInvocation), or the number of methods that store data in this variable (incomingInvocation). In both cases, the RBSS sorting strategy will consider the size of the list of methods as a criterion for evaluating the popularity of the variable.

- 2) Score attribution by category RBSS then rates the popularity of each item in the suggestion list based on the number of times other objects in the system use it. To assess the popularity score of items in a suggestion list partition, we assign a weight to each link that a given item can have. Thus, the total number of links of an element represents its popularity score, which is one factor that conditions the element's positioning (method, Class, Variable) in the suggestions list. for this, we use the RBSSorter algorithm Fig. 3. This algorithm uses algorithms from Fig. 4 and Fig. 5. The position of an element in the completion list considers the context of the word to be completed. So, we will contextualize the sorting of the list according to the type and the current situation. For example, in the case of a list having methods, variables, classes etc. as a completion proposal, the RBSS strategy after having sorted this list will first consider the context before displaying the result. If we are in a method context, RBSS will place the most popular methods at the top of the list and then complete the list with other popular elements (variable, class, ...) from this list and thus vice versa.
- 3) Recombination of the elements (see Fig. 6) of the partitions into a single list. For this we rely on the following RBSS Recombiner algorithm:

```

SORT
INPUT: L, the list of Candidates
RETURN: L', the sorted list of candidates
LOCAL: chg, the list of couple (l,c) where l in L and c
the computed score
For each l in L
  Chg add (l, score(l))
End for
For i=1 to threshold do
  For each (l,c) in chg
    (l,c) ← (l, (c + SUM Score(l') l' In Neighbor(l) ) /
              1+ (c + SUM Score(l') l' In Neighbor(l) ))
  End For
End for
Sort chg
RETURN first chg
RETURN seq
END.

```

Fig. 3. Algorithm RBSSorter: Sort the List of Candidate According to their Score

```

NEIGHBOURG
INPUT: l, a Candidate
RETURN: N, a list of neighbor of l
Case type of l :
  Method : N ← senders(l)
  Class : N ← AllRefInside (l)
  Variable : N ← OutGoingInvocation (l)
End Case
RETURN N
END.

```

Fig. 4. Algorithm Neighbour: Find Neighbor of a Candidate

III. EXPERIMENTS AND DISCUSSIONS

In a recent study [2] Hellendoorn and others present a case study on 15,000 code completions that were applied by 66 real developers. They find that many aspects of real-world completions are not represented in synthetic benchmarks and tested completion tools were far less accurate on real-world data. Worse, on the few completions that consumed most of the developers' time, prediction accuracy was less than 20% – an effect that is invisible in synthetic benchmarks. For these reasons we choose to test our strategy on real-world system.

```

SCORE
INPUT: l, a Candidate
RETURN: c, the score for l
c ← #Neighbour(l) / 1+ #Neighbour(l)
RETURN c
END.

```

Fig. 5. Algorithm Score: Give a Score to a Candidate

```
RECOMBINER
INPUT : List of candidates list sorted by type (L')
OUTPUT : List of candidates(L)
BEGIN
  L' := empty list
  For i = 1 to lenght(L) do :
    if type-elt-context = type-elt (L[i]) then
      Add elements of L at the beginning of L'
    else
      Add elements of L at the end of L'
    EndIf
  EndFor
RETURN (L')
END
```

Fig. 6. Algorithm RBSSRecombiner: Recombine Partitioned Lists.

A. Setup

For the evaluation the data used is as follows:

- We tested a large body of 1000 tokens in the pharo 9.0 source code to clearly show the impact. Pharo is a pure object programming language in which everything is object. the criteria that guided this choice are : the length of the object's source code, the type of the object (instance side, class side, ...), for the evaluation. Finally, we compare the score of our model with some basic strategies. The main objective of this evaluation is to empirically answer the research question about RBSS:
- How accurate is the RBSS method? We use this dataset as the basis for our assessments. We use the context and scope of each source code element.

Table 1 shows the data statistics, where LOCs are the line of codes, Files are the number of files, and Total Projects are the number of included projects.

TABLE I. DATA STATISTICS

Packages	Object Types	LOCs	Tokens
Epicea, Collections-Strings, Colors	Instance side	1119	7514
Epicea, Collections-Strings, Colors	Class side	370	2860
Total		1489	10374

We have tried different scope granularities such as class scope, method scope, and block scope. To the test pattern, we used 10,374 Entries names range in length from 3 to 12, with an average of 5. Accuracy assessment measures the difference between the expected result and the result obtained. We are looking for the right word to be placed in top positions at best. Otherwise, it occupies the highest position in the list. For this purpose, we will look at the top2, top 3, top4 and top5. We stop at this level because [1] have shown that beyond position 5 the programmer continues to type. By the way the words are not usually long in Pharo. Experimental result We evaluate the accuracy of our completion models according to standard metrics. We mainly consider top-K accuracies implying that the correct completion was often near the top

of the suggestion list. Meaning that, we evaluated the RBSS strategy according to the position of the expected word in the list of suggestions. We did not dwell on the speed of execution, which will be the subject of another study. For this, we intend to exploit optimization techniques such as indexing and dynamic programming. Table 2 shows the results obtained within the Pharo environments.

Through empirical evaluation of RBSS in both environments, we were capable to show its ability to improve the accuracy of the list of returned candidates. The tests results carried out according to the defined case show that with the RBSS strategy, If we take the first five candidates from the list, in 61,6% of cases, we have the right candidate in the list. From this point of view, the RBSS strategy is improved on auto-completion. In conclusion, we have shown that the sorting strategy based on the popularity of the elements improves the precision of the code completion system, that is, the positioning of the elements in the suggestion list. However, although RBSS performs better, we did not consider the execution speed aspect, which we intend to improve in future studies.

IV. THREATS TO VALIDITY

As any empirical evaluation, the results of our experiments are subject to threats to validity. We identified the following noteworthy threats:

- The studied system might not entirely represent a larger population of systems, either from another application domain or written in another programming language. This is always a complex threat to mitigate as there is little information on what property of a system is essential to ensure representativeness. Replication of the experiment for other systems must be realized. This said, we strongly believe our approach is independent of the programming language and the application domain.
- The way in which we setup our experimentation may introduce bias. We also believe Pharo and visual works are credible, real world, non-trivial, case study. It was medium to big system and it includes a significant number of completion tools and options. However, we firmly believe that our approach is language independent.
- We tried with different type of object (instance side, class side, traits). This was done to eliminate a possible problem with the obviously simple solution working for any kind of object.
- Internal threats to validity are related to the implementation of our approach. It is still possible that our approach implementation contains errors that can affect our results' exactitude. We manually studied a subset of the results to counter this threat and did not find any obvious errors. Bias concerning developer working habits might also occur in our selection of evaluation subjects. We selected various packages from the two studied systems to reduce this risk, all issued from different areas. Thus, we believe the objects represent a heterogeneous enough population of the source code element.

TABLE II. DATA COMPARISON

Token Type	Method			Class			Variable		
	Top3	Top4	Top5	Top3	Top4	Top5	Top3	Top4	Top5
RBSS	56.4%	59.7%	61.6%	50.4%	55.8%	63.4%	38.4%	39.9%	47.6%
Pharo	55.7%	59.6%	61.5%	50.7%	55.5%	63.4%	37.7%	39.8%	47.6%

V. SOME WORK ON CODE COMPLETION

Code completion. With the birth of IDEs, code completion research has received much attention in recent decades. In [8] authors present a largescale study of user interactions with auto-completion. They found that lowerranked auto-completion suggestions receive substantially lower engagement than those higherranked. They note that users are most likely to engage with auto-completion after typing about half of the query, and in particular at word boundaries. They also found that the likelihood of using auto-completion varies with the distance of query characters on the keyboard. In a first study, the traditional models which are based on the formal structure of the programs, that is to say on the syntactic information and the static properties of the code. These syntactic approaches have been the most explored [9], [4], [7]. In the paper [12] we can read: Software engineering and programming languages (SE / PL) should make the same transition as research on natural language processing, assisting traditional methods that only take into account the formal structure of the programs, that is to say the information on the statistical properties of the code, and also exploiting repetitive and predictable elements of the source code. Essentially, three completion techniques are recited, each using the information in the example database differently. It is about : A Frequency Based Code Completion System (FreqCCS) uses the frequency of method calls to decide their suitability and suggests the most frequently used method. An Association Rule Based Code Completion (ArCCS) which is a statistical learning technique for finding interesting associations between elements in data, ArCCS exploits the rules $m \rightarrow n$ which, if the method m is used, the method n is frequently called and suggested. The Best Matching Neighbors code completion (BMN) which is the modification of the K-Nearest-Neighbors machine learning algorithm, BMN adapts the KNN to suggest a variable v . The probabilistic models or statistical language models (Statistic Language Model). Recent work has started to examine linguistic models based on statistical learning [17], [18], [10] aiming to model the source code as statistical language learning models. These approaches offer an exciting new goal of the code completion problem that suggestions can capture the deeper meaning of terms' semantic and idiomatic meaning. These are, among others : N-gram models and Recurrent Neural Networks (RNN). The n-gram models which exploit probabilistic models and predict each token based on the probability of the preceding token. To deal with data scarcity, an N-gram data model estimates the probability of a sentence by modeling language as a Markov chain of order. The probability of the next word in the sentence (phrase) depends only on the previous words [11]. The Recurrent Neural Network outperforms the n-gram and predict each token (node to predict) sequentially. For example [16] proposes an approach that combines RNNs with networks of pointers to complete the code. In [13] authors explore the use of neural network techniques to automatically

learn code completion from a large corpus of dynamically typed JavaScript code. Authors propose a neural network model and believe that neural network techniques can play a transformative role in helping software developers manage the growing complexity of software systems. Performance measurement of these approaches. The following metrics are the most used [15]: precision , recall and F-Measure. Whose formulas are:

$$precision(P) = \frac{Recommendations_{made \cap relevant}}{Recommendations_{made}}$$

$$recall(R) = \frac{Recommendations_{made \cap relevant}}{Recommendations_{relevant}}$$

$$F - measure = \frac{(2 * P * R)}{(P + R)}$$

The measurement F is called the harmonic mean of recall and precision. Usually, it is difficult to achieve optimal results simultaneously for recall and precision. For example, if all words are classified as irrelevant, the resulting recall score will be 100% where the accuracy score will be low. Therefore, measurement F is a compromise between recall and precision. The score range for measure F is 0 to 1 ; the higher score implies a better classification model.

VI. CONCLUSION

In this paper, we have proposed an approach for improving code completion. This approach if used increase the productivity of coders. In addition, this mechanism offers advantages in terms of reducing typing errors and time while increasing the efficiency and productivity of programmers. We have examined our approach with Pharo a dynamic object language. The experimental results have shown that our proposed method surpasses existing methods in terms of effectiveness and efficiency. It turns out that our sorting strategy significantly improves code completion. The accuracy of the list of suggestions (the right candidate is in the top 3 first candidates 59.6% of the time) compared to the Pharo 9.0 completion engine sorting.

Several improvements and perspectives are possible despite achieving an adequate precision sorting strategy. Indeed, the RBSS sorting strategy proposed in this work is limited and has shortcomings. The essential concerns its execution speed : RBSS is slower than the alphabetical sorting strategy, i.e. the strategy takes much longer to display list of suggestions. Thus, the main perspective considered concerns optimizing the RBSS sorting strategy to consume less resources. This optimization requires for example the creation and initialization of an index or of an array in which RBSS will read and that the array is updated automatically at a precise time or at a given

frequency to go fast. And this takes into account the size and the appropriate structure of this table. We also plan to compare the performance of our system with the latest systems proposed in the literature. Another perspective is to evaluate our approach on a large set of developers' community to collect their comments and quantify the pros and cons of our approach. We also wish to extend and test our approach on other languages and environments because currently, our results are valid than Pharo.

REFERENCES

- [1] Jin, X., & Servant, F. (2018). The Hidden Cost of Code Completion: Understanding the Impact of the Recommendation-list Length on its Efficiency. Virginia Tech.
- [2] V. J. Hellendoorn, S. Proksch, H. C. Gall and A. Bacchelli, "When Code Completion Fails: A Case Study on Real-World Completions," 2019 IEEE/ACM 41st International Conference on Software Engineering (ICSE), 2019, pp. 960-970, doi: 10.1109/ICSE.2019.00101.
- [3] Murphy, M. K. (2006). How are java software developers using the eclipse IDE ? IEEE Software, 23(4), pp. 76–83.
- [4] Robbes, R. a. (2008.). How program history can improve code completion. Proceedings of the 2008 23rd IEEE/ACM International Conference on Automated Software Engineering.
- [5] Hayatou, O., Kolyang, & Moulla, K. (2019). Sur l'utilisation des mesures de qualité lors des phases de développement dans l'industrie logicielle camerounaise. Maroua: Université de Maroua.
- [6] ZAKRANI, A. (janvier 2020). Estimation des coûts de développement de logiciels par un réseaneuronal RBF flou.
- [7] D. Hou and D. M. Pletcher. (2011, Septembre). An evaluation of the strategies of sorting, filtering, and grouping API methods for Code Completion. ICSM '11: proceedings of the 2011 27th IEEE International Conference on Software Maintenance, 3-6.
- [8] Mitra, Bhaskar et al. "On user interactions with query auto-completion." Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval (2014).
- [9] Han, S. D. (2009). Code completion from abbreviated input. Automated Software Engineering, 2009. ASE'09. 24th IEEE/ACM International Conference on. IEEE.
- [10] Bielik, P. V. (2016). PHOG: probabilistic model for code.
- [11] Raychev, V. P. (2016). Probabilistic model for code with decision trees.
- [12] Allamanis, M. e. (2018). A Survey of Machine Learning for Big Code and Naturalness.
- [13] Chang Liu, X. W. (s.d.). NEURAL CODE COMPLETION. University of California, Berkeley.
- [14] Amann, S. e. (2016). Une étude de l'utilisation des studios visuels dans la pratique. 23e Conférence internationale IEEE sur l'analyse, l'évolution et la réingénierie logicielles (SANER), 1.
- [15] Rahman, M. W. (2020). A Neural Network Based Intelligent Support Model for Program Code Completion. Scientific Programming.
- [16] Jian Li, Y. W. (2016). Code Completion with Neural Attention and Pointer Networks.
- [17] Hu, S. X. (2019). Scope-aware code completion with discriminative modeling. Journal of Information Processing, 27, 469-478. .
- [18] Liu, F. L. (2020). A self-attentional neural architecture for code completion with multi-task learning. Proceedings of the 28th International Conference on Program Comprehension, 37-47.

Revisiting Polyglot Persistence: From Principles to Practice

Omar Lajam, Salahadin Mohammed
Information and Computer Science Department
King Fahd University of Petroleum and Minerals
Dhahran, Saudi Arabia

Abstract—To cope with the rapid advancements in information technologies, many database systems have been developed in the last decade to satisfy various data storage requirements, such as NoSQL databases. In many cases, using a single database system cannot be an option because of the limitations posed on the functionalities of the software application. Therefore, applications may use multiple distributed storage databases that complement each other to satisfy the conflicting requirements. Such applications that are called polyglot persistent applications. However, the practical implementation of polyglot persistence and its complexities have not been studied enough. In this paper, the most recent studies related to polyglot persistence are reviewed. Database systems are classified based on their data storage model, and their use cases are discussed. The principles of polyglot persistence and its challenges are expounded. The implementation architectures of polyglot persistence applications are categorized into Application-coordinated Polyglot Persistence, Service-oriented Polyglot Persistence, Polyglot- Persistence-as-a-Service, and Multi-models Databases. An analysis of the issues related to each architecture is presented. In light of the study findings, a practical polyglot persistence implantation strategy is proposed. The outcomes of this work can help design future polyglot persistence applications and influence future research on how to resolve the complexity involved in polyglot persistence solutions.

Keywords—Database system; database architecture; relational database; NoSQL; distributed storage; multi-model database; review; classification

I. INTRODUCTION

Data stores have been an integral component of software applications since the emergence of information technology systems, including shopping, accounting, medicine, and games applications. There is a wide range of database systems that are available for storing data, such as MySQL, MongoDB, and Cassandra, to name a few. These database systems are usually classified based on how they model and store the data. Nevertheless, relational databases are the most popular databases used in the industry [1], and they are the default option for typical applications.

Despite their popularity, relational databases have some limitations, such as expensive queries and vertical scalability, that make the selection of non-relational databases vital. As the application gets popular and the database gets burdened with millions of records, there becomes a need for using more flexible databases, such as NoSQL databases, to support features not supported by relational databases. At the same time, the complete abandonment of relational databases cannot be an option in many cases because they also support important

features not supported by other databases, like data consistency and transactions atomicity. To avoid sacrificing any of the different database features, the need for using more than one database system within an application to fulfill the conflicting requirements raises.

When an application uses more than one database system, it is called a polyglot persistence application. Taking this decision of having a polyglot persistence environment is not straightforward because it increases programming complexity and requires developers' knowledge of different database systems. However, successful implementation of polyglot persistence has the great advantage of having the different database systems complement each other and satisfying the conflicting requirements.

In order to ease the development of polyglot persistence applications, several architectures are proposed in the literature on how to design and implement them. Nevertheless, the following problems are identified by this study on those proposals: 1) There is no clear distinction or categorization for the different proposed architectures, which make them irrelevant to each other and difficult to be compared. 2) There is no detailed discussion on the challenges introduced by polyglot persistence architectures, while the focus is mostly on their advantages. 3) Most of the proposed architectures are not abstract in the sense that they cannot be applied on any application domain, and they are mostly targeting specific domains (e.g., e-commerce and healthcare), which makes them, in many cases, lack generalizability.

This review study aims to address these problems and build new knowledge based on the literature findings. The main contributions of this work can be highlighted as follows:

- Review for the most recent studies related to polyglot persistence.
- Classification for database systems with a detailed discussion on their characteristics.
- Description of polyglot persistence and its principles.
- Categorization for the architectures in which polyglot persistence applications can be implemented.
- Analysis of the problems associated with each polyglot persistence architecture.
- A practical strategy for polyglot persistence implementation.

To the best of our knowledge, this is the first work that expounds polyglot persistence comprehensively, from principles to practice. This work opens new opportunities for future research to address polyglot persistence challenges. It guides future polyglot persistence research towards more mindful solutions that consider the theoretical and practical aspects of polyglot persistence. This work can also help practitioners make wiser design decisions when building polyglot persistence applications. In addition, it may serve as a reference for understanding database systems concepts and challenges.

The rest of the paper is organized as follows. Section II discusses the related work. The methodology of this work is explained in Section III. Section IV presents database systems classification. In Section V, an explanation for polyglot persistence principles is given. Section VI describes polyglot persistence architectures and challenges. Section VII outlines the proposed polyglot persistence implementation strategy. And finally, Section VIII concludes the paper.

II. RELATED WORK

Few studies have discussed polyglot persistence concepts, architectures, and challenges. Gessert and Ritter [2] were the first who classified polyglot persistence based on research and industry into three patterns: application-coordinated polyglot persistence, microservices, and polyglot database services. The authors then described them again in [3], where they gave brief details about each pattern without mentioning issues related to each of them.

Another attempt to classify polyglot persistence is given by Khine and Wang [4] based on the polyglot persistence solution orientation. They classify the polyglot persistence solutions into three types: domain-oriented solution, query language-based solution, and other solutions (e.g., frameworks, middleware, and multi-model databases). A main observation on their classification is that it lacks disjointedness and holism.

Wiese [5] discussed polyglot databases architectures and challenges. Three architectures are described: polyglot persistence, lambda architecture, and multi-model databases. However, the lambda architecture is part of polyglot processing, not polyglot databases, as presented in [6].

Jaroslav [7] demonstrated some possible strategies for building an infrastructure that operates on integrated SQL and NoSQL databases. The study provides some approaches to construct such integrated database architectures, mainly by using multi-model databases and multi-level modeling, where interactions occur within and between at least two levels of connected databases.

Clearly, polyglot persistence is not studied enough. In this work, we try to study polyglot persistence principles, architectures, challenges, and implementation, based on the literature findings, as comprehensively as possible.

III. METHODOLOGY

The main objectives of this study are to explore how polyglot persistence applications can be architected and to understand polyglot persistence advantages and challenges. The following methodology is used in order to accomplish the study objectives.

A. Literature Review

This work is mainly a review that surveys the literature to gain knowledge on the topic. Four databases were searched to extract the related studies: IEEEExplore, ACM Digital Library, Web of Science, and Google Scholar. These four databases were chosen because they can capture most related studies. The search string used was 'polyglot persistence'. The inclusion criterion was to include any study that proposes a polyglot persistence architecture, model, or framework. The resulted studies were inspected, and 18 relevant studies were included. The selected studies were then downloaded and fully read. Each study was summarized, and all results were aggregated into an Excel file in a tabular format.

B. Problem Identification

At this stage, several problems were identified in the reviewed studies. First, many different architectures lack a clear distinction or categorization, making them unrelated to each other and difficult to compare. Second, there is no comprehensive treatment of the issues posed by polyglot persistence architectures, with the emphasis being placed mostly on their benefits. Third, most proposed architectures were specific to a few application domains, e.g., e-commerce and healthcare, which make them difficult to be generalized for other domains. In other words, many proposed architectures were developed with specific database requirements in mind.

These three problems are consistent with what Khine and Wang have found [4], where they stated that it has not yet been determined which architectures are best suited for different kinds of applications and how polyglot persistence can be implemented. Additionally, they observed that the benefits and limitations of polyglot persistence are still open research topics for academics and professionals.

C. Classification and Analysis

After identifying the problems, an intensive investigation was carried on to identify polyglot persistence principles, architectures, challenges, and implementation. Search databases were searched again with the same string. The relevant studies were analyzed, and more information was gathered from the literature to build knowledge that addresses the identified problems.

IV. DATABASES CLASSIFICATION

There are different options for storing application data, ranging from simply being stored in a file to being stored in a sophisticated data storage system, depending on the degree of complexity of the application requirements. Database Management Systems (DBMSs) are database systems that manage data storage and retrieval. The implementation of a DBMS specifies how the data will be structured and stored into the disk, how the queries will be processed, how access will be granted, and many other functions. What distinguishes one DBMS from another is its functions and features. DBMSs can be categorized into relational (RDBMS), referred to as SQL databases, and non-relational or NoSQL (Not only SQL) databases.

A. Relational Databases

Relational databases use the relational data model to store the data. It is the most popular model for storing structured data. It was first introduced by E. F. Codd in 1970 [8][9]. It is very mature, stable, trusted, and well researched. In this model, the data is organized as tables, and these tables can have relationships with each other. Examples of relational databases include MySQL, PostgreSQL, and Oracle.

Relational databases support transactions that obey ACID properties (Atomicity, Consistency, Isolation, and Durability) [10]. The atomicity property ensures that all instructions of a transaction will be executed at once, i.e., a transaction is an atomic unit of processing. The consistency property guarantees that the database state is always consistent, and the correct execution of a transaction takes the database from one consistent state to another. The isolation property makes each transaction completely isolated from others, and its effect on the database does not become visible until it is committed. This noninterference transaction execution can be achieved using concurrency control techniques [11]. The durability property, also referred to as permanency, ensures that changes made by a successful transaction will not be lost by subsequent unsuccessful transactions [12].

Relational databases have fixed database schema. They enforce data consistency and integrity. They store the data efficiently with minimal redundancy and maximal space utilization [13]. They have powerful query language. And lastly, they have a great community that provides support and help.

Relational databases manifest some drawbacks under some situations, especially when the number of database users dramatically increases and when the data volume becomes too huge. That is mainly due to the nontrivial processing required for user queries and the difficulty of operating on a distributed architecture. With massive amount of data, the relational database requires powerful machine to operate efficiently. The only option to scale the database system up is to upgrade the machine to a more powerful one. In other words, relational databases are only vertically scalable. Because they usually run on one machine, relational databases are prone to the single point of failure threat.

Relational databases are not suitable for unstructured, semi-structured, and graph data. They are not suitable for applications that store schema-less or schema-free data. They incur a high cost for complex query processing due to the table joins and constraints checking involved. They are less suitable for high-velocity ingestion due to the schema constraints validations. The relational database infrastructure (i.e., server machine) cost is expensive due to the powerful processing and storage space resources it needs, especially when the number of simultaneous users and/or the data volume becomes huge [14].

B. NoSQL Databases

The term "NoSQL" can be interpreted as "not using SQL query language", or can be interpreted as "Not only SQL", where the latter implies either the support of a database system for a query language that is similar to SQL or implies the co-existing of a non-SQL database with a SQL database in a

common polyglot persistence environment. There is no agreed-upon definition of what "NoSQL" is stand for [15].

The main characteristic of NoSQL databases is their ability to operate in a distributed architecture, running on a cluster of commodity hardware. In NoSQL databases, there is almost no referential integrity constraint among data objects. Therefore, processing data residing in many different machines is feasible, and horizontal scalability is enabled by simply adding new processing and storage resources without replacing old ones. In addition, distributed storage architecture enables the migration of processes to data and data to processes, which facilitates big data analysis tasks.

An important problem with SQL databases solved by NoSQL databases is the data structure *impedance mismatch* [16]. The in-memory data can be kept in complex structure (e.g., nested lists), while with SQL databases, the data is always in a simple tabular format. This difference between the two stores (in-memory and database) causes the impedance mismatch and requires translation work upon data writing and reading to and from the database. For object-oriented programming language, it would be more favorable to replicate the data objects stored in memory directly into the database. For SQL databases, this problem is mitigated by the Object-Relational Mappers (ORMs) [17], where they take the responsibility to map data objects to their corresponding underlying database structure. With most NoSQL databases, the in-memory data structure can be stored as-is into the database, and this feature reduces programming overhead and enhances the performance.

NoSQL databases are schema-less, and they do not enforce data integrity constraints. That makes NoSQL databases more efficient because constraints checking and integrity validation upon data insertion are eliminated [18]. The distribution architecture of NoSQL databases makes them fault-tolerant because they are not prone to the single point of failure threat. Data replication across distributed storage nodes in NoSQL databases is easy because there is no obligation to the ACID properties. Alternatively, NoSQL databases adhere to the BASE properties (Basic Availability, Soft state, Eventual consistency) [12]. The basic availability property ensures that every request will get a response. However, consistency among responses is not guaranteed, and multiple users requesting the same data object can get different versions. The soft state property allows the database system to remain inconsistent after query execution. The eventual consistency property promises to propagate the changes to storage nodes until eventually the entire distributed database system becomes on a global consistent state [19].

NoSQL databases are considered "non-relational" databases because their models are divergent from the traditional relational data model and implemented differently. NoSQL data models are categorized into Key-value, Document, Column Family, and Graph data models [20]. The next discussion for each model is mostly inspired from [15] [20], [21], and [22].

1) *Key-Value Model*: This is the simplest data model, where the data object is stored as a pair consisting of a key and a value. The key is a unique alpha-numeric identifier for the value. The value can be a string or complex lists and sets, with no constraints on its content structure. The structure of this

model is very similar to hash tables and dictionaries. In this model, the data can only be searched by key, i.e., the value is not searchable. Examples of key-value databases include Redis and Memcached.

The simplicity of this model makes it scalable and suitable for application that requires fast access to self-contained schema-less data. Examples of these data are user profiles, web sessions, shopping carts, and products information. On the other hand, the key-value model is not suitable if relationships exist between data objects, data is queried by its value, values are updated frequently, or for operating on multi-key transactions.

2) *Document Model*: This model can be considered an expansion to the key-value model, where the value contains semi-structured data and can be fully searched and indexed. Each data object is stored in a document that contains one or more keys. Groups of logically related documents are called collections, which are equivalent to tables in relational databases. To get the flexibility in accessing the data by its value, the database may store metadata that describes the allowable value structure and types. The database can retrieve part of the document based on the user query. The document can be formatted in a standard data exchange format such as XML, YAML, JSON, or BSON (Binary JSON). Examples of document databases include MongoDB and Couchbase.

The design of this model is inspired by a business software called Lotus Notes [23], a document database that enables sharing data across a local network [24]. Document databases use cases include storing and managing large-size collections of text files, such as literal documents, email messages, and XML files. Also, aggregated data objects such as products information or user profiles which are accessed at once together, are another use case for document databases. In general, document databases are best used for searchable data that has no fixed schema and which may add many nulls in an equivalent relational database. Document databases are not the best option for complex application queries or for transactions that require accessing multiple documents at once.

3) *Column Family Model*: Column family (or wide-column) model stores data objects in key-value pairs, where the value points to a second-level of key-value pairs. These second-level keys are called columns, and a subset of them forms a column family. The values can be accessed by any key in the first or second level.

One of the first column databases is BigTable [25], where it was designed to handle big data on a petabyte-scale. Another example of a column-family database is Cassandra [26], with a slightly different design philosophy that supports nested columns.

Column family databases may be the best choice with structured data when the distributed architecture is used, with data batch processing on a large scale, or real-time distributed big-data analysis tools such as MapReduce [27].

4) *Graph Model*: The graph model stores the data object as a graph consisting of nodes and connection edges. The nodes represent data objects while the edges represent relationships between them. Relationships are associated with properties, and two nodes can have one or more relationships. This model

is schema-less, and nodes and edges can be inserted with any content. Examples of graph databases include Neo4j and JanusGraph.

This model is the only NoSQL model that supports relationships and ACID transactions. In fact, this model is closer to the relational model but categorized as NoSQL because of its dissimilarity with the relational model in how the data is structured and queried [28]. A key difference between the graph model and the relational model is in the query cost, where the navigation along the graph network to explore information is cheaper with graph database due to the absence of the expensive join operations. Another obvious difference is that the SQL query language is not supported in graph databases [28].

The graph representation of the data helps extract information that is hard to get with other models. The data of real-world problems that have interconnected entities, such as social networking, maps, products recommendations, pattern detection, network topologies, or any problem that can be represented as a graph, is a good candidate to be stored in a graph database. Nevertheless, graph databases are not good in horizontal scalability and big data processing.

V. POLYGLOT PERSISTENCE PRINCIPLES

Due to the availability of many heterogeneous database systems, the decision of which database system should be used for a given application can be embarrassing. The concepts of polyglot persistence can be utilized in such cases. This section explains the meaning of polyglot persistence and spot the situations in which it is really needed.

A. What is Polyglot Persistence?

The term polyglot persistence was first coined in an online blog by Scott Leberknight in 2008 [29], and it then became famous after the book [15]. Leberknight explained the meaning of polyglot persistence by "*like polyglot programming, is all about choosing the right persistence option for the task at hand*". The term 'polyglot' implies the ability to talk to more than one database system. Polyglot persistence can be defined as *a situation in which different parts of data are stored in the most persistent database system that satisfies the storage requirements*.

Traditionally, relational databases were the default acceptable persistent option for data storage. However, the appearance of non-relational databases has changed the norm since there can be non-relational databases that are more persistent in some cases. The determination of the most persistent database system is totally dependent on the application's storage requirements.

A common example to illustrate the meaning of polyglot persistence is with an e-commerce application. In a typical e-commerce application, queries about clients' shopping data can be easily answered using a key-value NoSQL database. However, if the interest is on what the client's friends have purchased, the problem becomes entirely different. To answer this question, a graph database should be used [15]. Fig. 1 shows an example of a possible implementation for polyglot persistence in an e-commerce application.

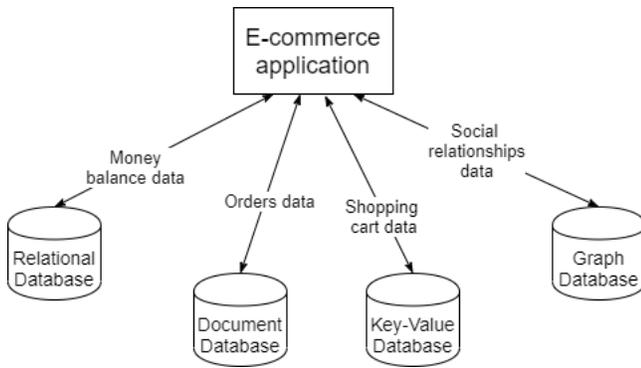


Fig. 1. Example of implementation for Polyglot Persistence in E-Commerce Application

Large and well-known applications are found to be employing polyglot persistence. Examples of these applications are Google, Facebook, Amazon, and Twitter. The practical aspects of polyglot persistence are not new for the industry, but they are not studied enough by the research community. There is a shortage of addressing the problems associated with polyglot persistence in the literature, especially with designing, implementing, and maintaining polyglot persistence architectures [3].

B. Why is Polyglot Persistence?

As shown in Section IV, each of SQL and NoSQL databases have their own characteristics and advantages. Polyglot persistence is employed to get the optimal benefits of each different database system. A usual concern is with database scalability, schema flexibility, data consistency, system availability, and performance [3].

In general, polyglot persistence is adopted to resolve conflicting requirements [3] [5]. These conflicts exist due to the limitations of the database systems, where no one database system can satisfy all requirements. With polyglot persistence, all requirements can be satisfied by using as many databases as needed. If there was a one-size-fits-all solution, as promised by NewSQL [30], then polyglot persistence can be overlooked.

Next, examples of three types of conflicting requirements: functional, non-functional, and data requirements, are discussed.

1) *Conflicting Functional Requirements:* The boundaries of database system functions can be determined by the available commands supported by its query language. Examples of these commands in SQL are SELECT and INSERT. The database functional requirements of an application can be determined by listing the commands it needs to perform on its data storage. A conflict in the functional requirements occurs when no single database system supports the entire set of query commands required by an application.

For example, consider a key-value database that only supports GET and PUT commands used by a simple web blog application. If the application added new features that require more complex queries, such as user authentication and online course registration features, then either the database system will be changed, or a new database system will be added

beside the existing one. The latter polyglot persistence solution eliminates the need for a complex data migration process from the legacy database into the new one [31]. Note that the conflict in functional requirements can also be related to the database system security commands, such as the commands related to user authentication and access control [32].

2) *Conflicting Non-Functional Requirements:* The CAP theorem states that three demanding non-functional requirements cannot be satisfied at the same time when designing an application on a distributed architecture: Availability, Consistency, and Partition tolerance [33] [34] [35]. Therefore, there must be a trade-off for these requirements when selecting the database system for a given application. To resolve the conflict, different parts of application data can be stored in different databases.

Many non-functional requirements are subject to the ability of the database system to operate in a distributed architecture on commodity hardware. This ability reduces hardware resource costs, increases fault tolerance and availability, raises processing power, enables data replication, and eases big data analysis [36]. Since NoSQL databases can be deployed in a distributed architecture, they can be used to satisfy the mentioned requirements. On the other hand, other non-functional requirements cannot be accomplished unless the database system runs on a single server. Examples of these requirements are data consistency and integrity.

3) *Conflicting Data Requirements:* The data requirements for an application can also encourage the decision of using more than one database system. For example, a customer profile data (e.g., name, age, job, etc.) may not be designed in a fixed schema since many details can be null-able, and they may be frequently changed along the lifetime of the application development cycle. In addition, it might be impossible sometimes to design a fixed schema because the shape of the data is unknown in advance, as with the case when the data is inserted based on the user preferences. An example of such data object is salary, which contains many data items like basic salary, insurance allowance, transportation allowance, etc. In these cases, the usage of a schema-less non-relational database is recommended.

On the other hand, there are cases where using a relational database is the only possible option. An example of such a case is with money balance data that is used, for example, to purchase products or services within a web application. In this case, ACID transactions must be used to ensure data integrity and avoid race conditions [37].

Another example to illustrate the conflicting data requirements is the speed of data write or read operations. Some data have higher priority for reading speed over writing speed, such as product information, while other data might have higher priority for writing speed over reading speed, such as viewers counter for a product. To satisfy these conflicting requirements, different database systems might be selected for each part of the data.

VI. POLYGLOT PERSISTENCE ARCHITECTURES

To implement polyglot persistence in an application, one can consider more than one architecture. According

to 18 reviewed studies that implement polyglot persistence, these architectures can be categorized into four categories: Application-coordinated Polyglot Persistence, Service-oriented Polyglot Persistence, Polyglot-Persistence-as-a-Service, and Multi-models Databases. Table I shows the four categories, the number of studies implemented each of them, and their references. A description for each of the categories is given in this section.

TABLE I. CATEGORIES OF POLYGLOT PERSISTENCE ARCHITECTURES

Category	Count	Reference
Application-coordinated Polyglot Persistence	6	[38], [21], [39], [40], [41], [42]
Service-oriented Polyglot Persistence	9	[22], [43], [44], [45], [46], [47], [48], [49], [50]
Polyglot-Persistence-as-a-Service	2	[51], [52]
Multi-models Databases	1	[53]

A. Application-Coordinated Polyglot Persistence

With this architecture, the application itself coordinates the polyglot persistence. This coordination requires the application to control the mapping of the data to databases, i.e., to have explicit knowledge about where each part of the data is stored. Typically, if the application is not small, it would be divided into modules [54] (aka packages or components). Each module is responsible for part of the application and has its own logic and functions. If the data managed by a module is specific to it (not shared by any other module), managing polyglot persistence would be simple because each module can have a different exclusive database system. However, usually, data application is shared by more than one module. In this case, many challenges to support polyglot persistence arise. Also, in some cases, a single module may have conflicting requirements that have to be satisfied using more than one database system.

To distinguish between these different cases and ease the discussion of the challenges of each of them, we classify the relationships between modules and databases within an application as follows:

- **One-to-One:** A module has a connection with one exclusive database.
- **One-to-Many:** A module has connections with more than one exclusive databases.
- **Many-to-One:** More than one module have connections with one mutual database.
- **Many-to-Many:** More than one module have connections with more than one mutual databases.

An application can have a combination of these relationships. The four relationships are shown in Figure 2, where 'M' stands for module and 'D' for database. The assumption is that different databases in the figure are of different storage models. Next, each of these relationships is discussed.

1) *One-to-One Relationship:* This is the simplest relationship. The module controls everything related to its data in one sole database. One query language can be used to manipulate the data. Thus, programmers need to learn only one query language. The application development will also

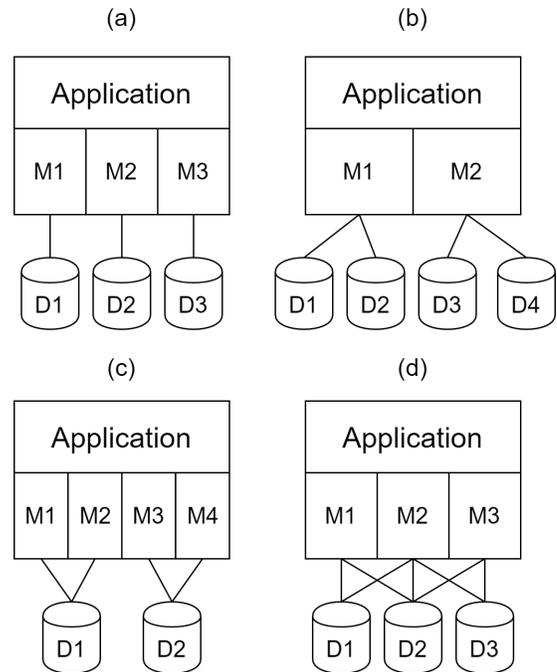


Fig. 2. Polyglot Relationships: (a) One-to-One, (b) One-to-Many, (c) Many-to-One, and (d) Many-to-Many

be easier because different programmers can work separately on different modules with knowledge about only one database system. A failure of one database will affect only part of the application and will not be propagated. Database configurations and security are taken care of by one module. A main concern here is with the design decision that will decide which database system is the best to satisfy the requirements for a given module. However, the approach proposed by [55] can ease the problem, where the functional, non-functional, and data requirements can be analyzed systemically to determine the most persistent database system.

2) *One-to-Many Relationship:* In this relationship, one module controls data stored by more than one database system, and that can cause several problems.

First, there will be a need to use more than one query language, one for each database system, which requires wider knowledge and longer training for developers. In addition, that may make the programming task more confusing. Possible mitigation to this problem can be by using a uniform query language for heterogeneous database systems [56], by a query mediator [57], or by translating SQL queries into NoSQL queries [58].

Second, cross-database consistency of dependent data objects stored in different databases needs to be managed by the module because there are no global referential integrity constraints enforced on the different databases. Consider Fig. 2(b). If dependency exists between a data object X on D1 and another data object Y on D2, then module M1 should maintain consistency across databases D1 and D2 by reflecting changes of X on Y and vice versa.

Third, running a query across different databases is not straightforward since data need to be processed and integrated

at the module, not at the database system, which may increase the query cost and require complicated data processing logic [59]. For example, a read query for a data object that is scattered on multiple databases may require several different sub-queries, one for each database, and the results of these sub-queries should be integrated and structured by the module. Data integration from different sources is studied in [60].

Forth, data redundancy might be a problem when different parts of the same data object are stored at different databases because common parts of the data object can be unnecessarily duplicated.

Fifth, in case of a failure of one database, this failure might be logically cascaded to other databases that are used by the same module in case a dependency exists between data objects.

3) *Many-to-One Relationship*: In this relationship, two or more modules use one mutual database. A main issue with this relationship is that the database configurations and security depend on more than one module, which may violate the Least Common Mechanism security design principle [61]. More on database security can be found in [62].

If all modules of the application are using the same database, then the polyglot persistence concepts are not applied. If this is not the case, then one can think of the modules that share the same database as one logical module, and the relationship becomes as if it were a one-to-one relationship. To simplify the control of a single database, a data manager (e.g., ORM) can be used as an intermediate layer between the modules and the database. It should control the database queries and configurations and mitigate security threats.

4) *Many-to-Many Relationship*: This relationship is the most complex, where each module uses at least two mutual databases. From the modules side, one can think of this relationship as a one-to-many. On the other hand, from the database side, one can think of this relationship as a many-to-one. Therefore, the same discussion of the two previous relationships can be said again here. However, the consistency problem is expended here because there will be a need to maintain cross-modules consistency for the data objects that are dependent on each other and stored in different databases, and are used by different modules. For example, in Fig. 2 (d), if a data object *X* on D1 is dependent on another data object *Y* on D2, and another data object *Z* on D3 is dependent on the same data object *Y* on D2, then we need to ensure consistency across modules M1 and M3 because they both share a common data object *Y* at D2.

A summary of the four relationships and their issues is given in Table II, where 'm' in the table header stands for 'many'. Note that the discussion here was at the module level, but it can be generalized to a larger programming unit, such as an entire application or even a set of applications, or smaller programming units, such as classes or methods.

B. Service-Oriented Polyglot Persistence

If the database is being used by more than one module or application, then the database can be *decoupled* from the application to reduce the complexity. In this case, only one *mediator* will be able to access and control the database. This mediator is an independent module or a small application that

TABLE II. ISSUES OF POLYGLOT PERSISTENCE RELATIONSHIPS

Issue	1-1	1-m	m-1	m-m
More than one query language might be needed	No	Yes	No	Yes
The module(s) need(s) to control cross-database consistency	No	Yes	No	Yes
Data integration might be required at the application side	No	Yes	No	Yes
The application should control cross-modules consistency	No	No	No	Yes
Data might be unnecessarily redundant	No	Yes	No	Yes
Database failure will be logically cascaded	No	Yes	No	Yes
Database configurations & security are dependent on more than one module	No	No	Yes	Yes

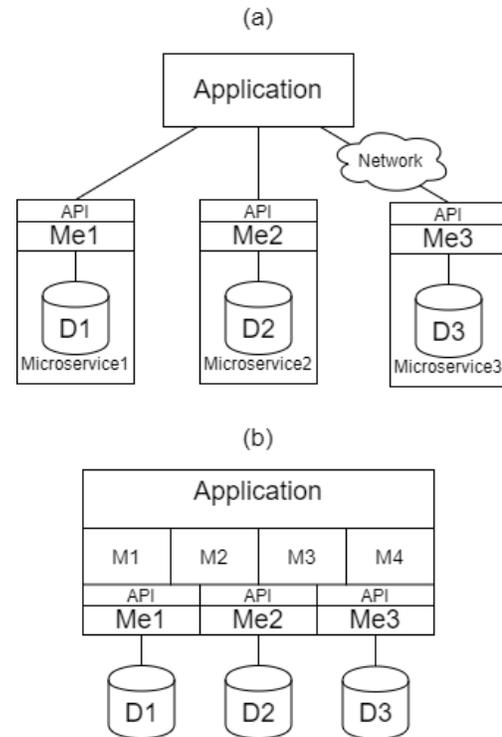


Fig. 3. Service-Oriented Architecture with (a) Microservices and (b) Modular Mediators

offers an API for external users. The degree of decoupling this mediator can have different levels. In one extreme, the mediator will completely be independent from the application and can be deployed in a different machine, and it can even be programmed with a different programming language. The API in this case will be network calls (e.g., REST API [63]). In this extreme, the mediator is part of what is called a *microservice* [64]. In the other extreme, the mediator is part of the application, and it offers an API as public function calls to other modules, or even other applications.

Regardless of the detailed structure of this architecture, it can be seen as service-oriented architecture [65], where the database is wrapped with a software controller (which is called inhere a mediator). Illustrative examples of this architecture are shown in Figure 3, where 'M' stands for module, 'Me' for mediator, and 'D' for database. In Fig. 3(a), the network cloud implies the possibility for the entire microservice to be remotely accessed.

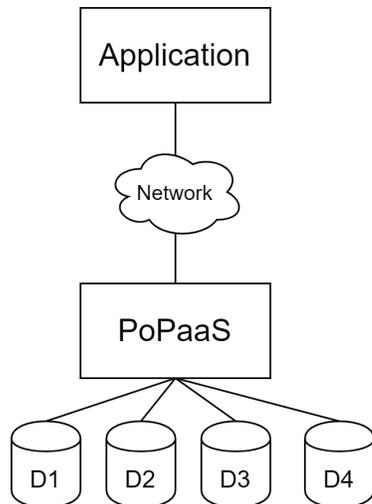


Fig. 4. An Illustrative Example for PoPaaS Architecture

A main advantage of this architecture is that different databases are accessed using similar APIs. The application is completely unaware about the underlying query languages of the databases. However, this architecture puts the load of managing consistency and cross-queries on the application modules. In case the relationships between the application and the databases are many-to-one or many-to-many, the same issues discussed in Section VI-A can be considered here.

Several strategies are proposed in the literature to manage polyglot persistence queries with microservices, and they are presented in [48].

C. Polyglot-Persistence-as-a-Service

In this architecture, the burden of managing polyglot persistence is conveyed to a completely different location, the cloud. The application here only provides the data requirements, and then these requirements should be satisfied by the Polyglot-Persistence-as-a-Service (PoPaaS) provider. The provider should automatically specify the appropriate database system for each segment of the data, based on its requirements, and then provide an API for data access. Such an API design strategy is proposed by [51].

The problem with this architecture is how the client can formulate the requirements in a standard format? Another problem is the selection of the appropriate database system for the given requirements, which should be automated based on quantifiable metrics [3]. A possible solution is to use an automated rule-based data model selection technique, as proposed by [52]. In reality, we do not know an example of such a service. An illustrative example of this architecture is shown in Fig. 4.

D. Multi-Model Databases

Instead of dealing with the complexity of managing multiple databases, one possible solution is to use a database that supports multiple data models, which is called a multi-model database. In this case, the application will manage a single database that fulfills its requirements. Multi-model database

engines can be designed to manage a combined data model that has the features of several data models. OrientDB and ArangoDB are two examples of such databases [66]. They support document, graph, and key-value data models in one database instance. OrientDB has a query language that is very similar to SQL, while ArangoDB has a new language called ArangoDB Query Language (AQL), which is similar to an extent to SQL.

A new paradigm to support more than one model is with the Flexible Schema Data Management (FSDM) [67]. This paradigm integrates the JSON data model into SQL databases. The stored JSON data is storable, indexable, and queryable, without the need for upfront schema definition. This support will reduce the use cases where polyglot persistence is needed because there will be no need to use NoSQL databases to store schema-less data. PostgreSQL and Oracle databases are examples of such database systems that support this feature [68].

Despite the support for multiple data models in those databases, they do not eliminate the demand for polyglot persistence because there are still non-functional requirements that are not satisfied, such as scalability and performance [69]. This emphasizes the fact that satisfying all polyglot persistence requirements in one database system is an engineering challenge [3], and apparently, having a one-size-fits-all solution is not yet feasible.

VII. POLYGLOT PERSISTENCE IMPLEMENTATION STRATEGY

In this section, we propose an implementation strategy that can aid the development of polyglot persistence applications.

A. Step 1: Database Requirements

The first thing to start with is the requirements. The correctness of the requirements should be ensured because the following steps will be dependent on them. Database requirements must be consistent with the application requirements, and they should include functional, non-functional, and data requirements. Database functional requirements should include the queries that will be used to manage application data. Examples of non-functional requirements include consistency, integrity, and availability. One of the most important parts of data requirements is the conceptual data model, which should be created at this step using Unified Modeling Language (UML) [70] or Entity-Relationship (ER) [71] diagrams, for example.

B. Step 2: Database Selection

Based on the gathered requirements, the database system should be selected. At this step, conflicts between database requirements should be identified and resolved by using as many database systems as needed. If there was no conflict, the implementation will normally proceed using one database system without considering polyglot persistence. Otherwise, different database systems should be selected carefully, considering the most persistent database system for each part of application data. A clear mapping between each part of the data and the selected database systems should be preserved. Note that steps 1 and 2 can be accomplished using the systematic approach proposed in [55].

C. Step 3: Database-Specific Data Models

Conceptual data models for each database system should be derived from the general conceptual data model identified in step 1, considering the process proposed by [72]. Since different database systems use different storage models, each of them may require a specific data model to determine database schema. At this step, using a database-independent schema declaration language can be helpful [73].

D. Step 4: Polyglot Persistence Architecture

In this step, the architecture in which the application will be implemented should be designed. The architecture determines how the application will be talking to the different selected databases. The architecture design should consider implementation cost, time, and resources. The discussion on polyglot persistence architectures given in Section VI can guide this step.

E. Step 5: Issues Identification

After selecting the architecture, polyglot persistence issues associated with the selected architecture should be identified. After identifying the architecture issues, a clear plan on how they will be resolved should be made. This plan should include the technical details on how each issue will be addressed. Again, the issues and their resolutions can be inspired by the discussion in Section VI.

F. Step 6: Application Development

This is the last step in which the actual implementation for the application and its infrastructure should start based on the results of the previous steps.

VIII. CONCLUSION

Large applications may require more than one database system to satisfy their requirements. This environment that operates on multiple databases is called a polyglot persistence environment. The polyglot persistence environment is fraught with many challenges and problems. This paper presented classification of database systems with details about their features. Polyglot persistence principles, its possible architectures, and the issues related to each architecture are identified. A polyglot persistence implementation strategy is proposed in light of the study outcomes.

The authors believe that this work has clarified most of the concepts related to polyglot persistence. This work can be a helpful reference for solving polyglot persistence problems.

Future research can further propose solutions to issues introduced by polyglot persistence. A protocol for inter-database negotiations might be devised to support interoperability between different database systems as a means to address issues related to polyglot persistence. Furthermore, future research can study how database functional, non-functional, and data requirements can be reported standardly. In addition, it can study the possibility of having an approach for representing and analyzing database requirements that can automatically determine the needed database systems.

ACKNOWLEDGMENT

The authors would like to thank King Fahd University of Petroleum and Minerals for the support and facilities provided to perform this research.

REFERENCES

- [1] (2022) Db-engines ranking. <https://db-engines.com/en/ranking> (accessed: 2022-03-02).
- [2] F. Gessert and N. Ritter, "Scalable data management: Nosql data stores in research and practice," in *2016 IEEE 32nd International Conference on Data Engineering (ICDE)*. IEEE, 2016, pp. 1420–1423.
- [3] F. Gessert, W. Wingerath, and N. Ritter, "Polyglot persistence in data management," in *Fast and Scalable Cloud Data Management*. Springer, 2020, pp. 149–174.
- [4] P. P. Khine and Z. Wang, "A review of polyglot persistence in the big data world," *Information*, vol. 10, no. 4, p. 141, 2019.
- [5] L. Wiese, "Polyglot database architectures= polyglot challenges." in *LWA*, 2015, pp. 422–426.
- [6] M. K. Bavirisetty. (2015) Polyglot processing - an introduction 1.0. <https://www.slideshare.net/MohanBavirisetty/polyglot-processing-an-introduction-10> (accessed: 2022-03-02).
- [7] J. Pokorný, "Integration of relational and nosql databases," *Vietnam Journal of Computer Science*, vol. 6, no. 04, pp. 389–405, 2019.
- [8] E. Codd, "A relational model of data for large relational databases," *Communications of the ACM*, vol. 13, no. 6, pp. 77–87, 1970.
- [9] E. F. Codd, *The relational model for database management: version 2*. Addison-Wesley Longman Publishing Co., Inc., 1990.
- [10] T. Haerder and A. Reuter, "Principles of transaction-oriented database recovery," *ACM computing surveys (CSUR)*, vol. 15, no. 4, pp. 287–317, 1983.
- [11] P. A. Bernstein, V. Hadzilacos, and N. Goodman, *Concurrency control and recovery in database systems*. Addison-wesley Reading, 1987, vol. 370.
- [12] N. Banothu, S. Bhukya, and K. V. Sharma, "Big-data: Acid versus base for database transactions," in *2016 International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT)*. IEEE, 2016, pp. 3704–3709.
- [13] J. L. Harrington, *Relational database design and implementation*. Morgan Kaufmann, 2016.
- [14] C. A. Györfi, D. V. Dumşec-Burescu, D. R. Zmaranda, R. Ş. Györfi, G. A. Gabor, and G. D. Pecherle, "Performance analysis of nosql and relational databases with couchdb and mysql for application's data storage," *Applied Sciences*, vol. 10, no. 23, p. 8524, 2020.
- [15] P. J. Sadalage and M. Fowler, *NoSQL distilled: a brief guide to the emerging world of polyglot persistence*. Pearson Education, 2013.
- [16] S. Ambler, *Agile database techniques: Effective strategies for the agile software developer*. John Wiley & Sons, 2012.
- [17] J. M. Barnes, "Object-relational mapping as a persistence mechanism for object-oriented applications," 2007.
- [18] B. Jose and S. Abraham, "Performance analysis of nosql and relational databases with mongodb and mysql," *Materials today: PROCEEDINGS*, vol. 24, pp. 2036–2043, 2020.
- [19] M. Diogo, B. Cabral, and J. Bernardino, "Consistency models of nosql databases," *Future Internet*, vol. 11, no. 2, p. 43, 2019.
- [20] A. Moniruzzaman and S. A. Hossain, "Nosql database: New era of databases for big data analytics-classification, characteristics and comparison," *International Journal of Database Theory and Application*, vol. 6, no. 4, 2013.
- [21] K. Srivastava and N. Shekoker, "A polyglot persistence approach for e-commerce business model," in *2016 International Conference on Information Science (ICIS)*. IEEE, 2016, pp. 7–11.
- [22] C. Shah, K. Srivastava, and N. Shekoker, "A novel polyglot data mapper for an e-commerce business model," in *2016 IEEE Conference on e-Learning, e-Management and e-Services (IC3e)*. IEEE, 2016, pp. 40–45.
- [23] K. Moore, "The lotus notes storage system," *ACM SIGMOD Record*, vol. 24, no. 2, pp. 427–428, 1995.

- [24] L. Kawell Jr, S. Beckhardt, T. Halvorsen, R. Ozzie, and I. Greif, "Replicated document management in a group communication system," in *Proceedings of the 1988 ACM conference on Computer-supported cooperative work*, 1988, p. 395.
- [25] F. Chang, J. Dean, S. Ghemawat, W. C. Hsieh, D. A. Wallach, M. Burrows, T. Chandra, A. Fikes, and R. E. Gruber, "Bigtable: A distributed storage system for structured data," *ACM Transactions on Computer Systems (TOCS)*, vol. 26, no. 2, pp. 1–26, 2008.
- [26] A. Lakshman and P. Malik, "Cassandra: a decentralized structured storage system," *ACM SIGOPS Operating Systems Review*, vol. 44, no. 2, pp. 35–40, 2010.
- [27] J. Dean and S. Ghemawat, "Mapreduce: simplified data processing on large clusters," *Communications of the ACM*, vol. 51, no. 1, pp. 107–113, 2008.
- [28] C. Vicknair, M. Macias, Z. Zhao, X. Nan, Y. Chen, and D. Wilkins, "A comparison of a graph database and a relational database: a data provenance perspective," in *Proceedings of the 48th annual Southeast regional conference*, 2010, pp. 1–6.
- [29] S. Leberknight. (2008) Polyglot persistence. [Http://www.sleberknight.com/blog/sleberkn/entry/polyglot_persistence](http://www.sleberknight.com/blog/sleberkn/entry/polyglot_persistence) (accessed: 2022-03-02).
- [30] A. Pavlo and M. Aslett, "What's really new with newsql?" *ACM Sigmod Record*, vol. 45, no. 2, pp. 45–55, 2016.
- [31] S. Velimeneti, "Data migration from legacy systems to modern database," 2016.
- [32] G. D. Samaraweera and J. M. Chang, "Security and privacy implications on database systems in big data era: a survey," *IEEE Transactions on Knowledge and Data Engineering*, vol. 33, no. 1, pp. 239–258, 2019.
- [33] S. Gilbert and N. Lynch, "Brewer's conjecture and the feasibility of consistent, available, partition-tolerant web services," *Acm Sigact News*, vol. 33, no. 2, pp. 51–59, 2002.
- [34] E. Brewer, "A certain freedom: thoughts on the cap theorem," in *Proceedings of the 29th ACM SIGACT-SIGOPS symposium on Principles of distributed computing*, 2010, pp. 335–335.
- [35] S. Gilbert and N. Lynch, "Perspectives on the cap theorem," *Computer*, vol. 45, no. 2, pp. 30–36, 2012.
- [36] S. Venkatraman, K. Fahd, S. Kaspi, and R. Venkatraman, "Sql versus nosql movement with big data analytics," *Int. J. Inform. Technol. Comput. Sci.*, vol. 8, pp. 59–66, 2016.
- [37] R. Paleari, D. Marrone, D. Bruschi, and M. Monga, "On race vulnerabilities in web applications," in *International Conference on Detection of Intrusions and Malware, and Vulnerability Assessment*. Springer, 2008, pp. 126–142.
- [38] C. Zdepski, T. A. Bini, S. N. Matos, and S. Hammoudi, "An approach for modeling polyglot persistence," in *ICEIS (1)*, 2018, pp. 120–126.
- [39] S. Prasad and M. N. Sha, "Nextgen data persistence pattern in healthcare: polyglot persistence," in *2013 Fourth International Conference on Computing, Communications and Networking Technologies (ICCCNT)*. IEEE, 2013, pp. 1–8.
- [40] A. M. C. de Araújo, V. C. Times, and M. U. da Silva, "Polyehr: A framework for polyglot persistence of the electronic health record," in *Proceedings of the International Conference on Internet Computing (ICOMP)*. The Steering Committee of The World Congress in Computer Science, Computer ..., 2016, p. 71.
- [41] S. Prasad and S. Avinash, "Application of polyglot persistence to enhance performance of the energy data management systems," in *2014 International Conference on Advances in Electronics Computers and Communications*. IEEE, 2014, pp. 1–6.
- [42] S. Nadkarni, A. Kadakia, and K. Shrivastava, "Providing scalability to data layer using a novel polyglot persistence approach," in *2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA)*. IEEE, 2018, pp. 1–5.
- [43] K. Trivedi, S. Shah, and K. Srivastava, "An efficient e-commerce design by implementing a novel data mapper for polyglot persistence," in *Advanced Computing Technologies and Applications*. Springer, 2020, pp. 149–156.
- [44] L. H. Z. Santana and R. dos Santos Mello, "A middleware for polyglot persistence of rdf data into nosql databases," in *2019 IEEE 20th International Conference on Information Reuse and Integration for Data Science (IRI)*. IEEE, 2019, pp. 237–244.
- [45] K. Kaur and R. Rani, "A smart polyglot solution for big data in healthcare," *IT Professional*, vol. 17, no. 6, pp. 48–55, 2015.
- [46] —, "Managing data in healthcare information systems: many models, one solution," *Computer*, vol. 48, no. 3, pp. 52–59, 2015.
- [47] R. Jiménez-Peris, M. Patiño-Martinez, I. Brondino, and V. Vianello, "Transactional processing for polyglot persistence," in *2016 30th International Conference on Advanced Information Networking and Applications Workshops (WAINA)*. IEEE, 2016, pp. 150–152.
- [48] L. H. Villaça, L. G. Azevedo, and F. Baião, "Query strategies on polyglot persistence in microservices," in *Proceedings of the 33rd Annual ACM Symposium on Applied Computing*, 2018, pp. 1725–1732.
- [49] H. Singhal, A. Saxena, N. Mittal, C. Dabas, and P. Kaur, "Polyglot persistence for microservices-based applications," *International Journal of Information Technologies and Systems Approach (IJITSA)*, vol. 14, no. 1, pp. 17–32, 2021.
- [50] L. G. Azevedo, R. d. S. Ferreira, V. T. d. Silva, M. de Baysar, E. F. d. S. Soares, and R. M. Thiago, "Geological data access on a polyglot database using a service architecture," in *Proceedings of the XIII Brazilian Symposium on Software Components, Architectures, and Reuse*, 2019, pp. 103–112.
- [51] F. Gessert, S. Friedrich, W. Wingerath, M. Schaarschmidt, and N. Ritter, "Towards a scalable and unified rest api for cloud data stores," in *GI-Jahrestagung*, 2014, pp. 723–734.
- [52] M. Schaarschmidt, F. Gessert, and N. Ritter, "Towards automated polyglot persistence," *Datenbanksysteme für Business, Technologie und Web (BTW 2015)*, 2015.
- [53] I. Košmerl, K. Rabuzin, and M. Šestak, "Multi-model databases-introducing polyglot persistence in the big data world," in *2020 43rd International Convention on Information, Communication and Electronic Technology (MIPRO)*. IEEE, 2020, pp. 1724–1729.
- [54] J. B. Dennis, "Modularity," in *Software Engineering*. Springer, 1975, pp. 128–182.
- [55] N. Roy-Hubara, P. Shoval, and A. Sturm, "Selecting databases for polyglot persistence applications," *Data & Knowledge Engineering*, vol. 137, p. 101950, 2022.
- [56] B. Kolev, P. Valduriez, C. Bondiombouy, R. Jiménez-Peris, R. Pau, and J. Pereira, "Cloudmdsql: querying heterogeneous cloud data stores with a common language," *Distributed and parallel databases*, vol. 34, no. 4, pp. 463–503, 2016.
- [57] R. Sellami and B. Defude, "Complex queries optimization and evaluation over relational and nosql data stores in cloud environments," *IEEE transactions on big data*, vol. 4, no. 2, pp. 217–230, 2017.
- [58] J. Rith, P. S. Lehmayr, and K. Meyer-Wegener, "Speaking in tongues: Sql access to nosql systems," in *Proceedings of the 29th Annual ACM Symposium on Applied Computing*, 2014, pp. 855–857.
- [59] G. C. Deka, "Nosql polyglot persistence," in *Advances in Computers*. Elsevier, 2018, vol. 109, pp. 357–390.
- [60] M. Lenzerini, "Data integration: A theoretical perspective," in *Proceedings of the twenty-first ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, 2002, pp. 233–246.
- [61] T. V. Benzel, C. E. Irvine, T. E. Levin, G. Bhaskara, T. D. Nguyen, and P. C. Clark, "Design principles for security," NAVAL POSTGRADUATE SCHOOL MONTEREY CA DEPT OF COMPUTER SCIENCE, Tech. Rep., 2005.
- [62] N. Ron Ben, *Implementing Database Security and Auditing*. Elsevier, 2005.
- [63] V. Surwase, "Rest api modeling languages-a developer's perspective," *Int. J. Sci. Technol. Eng.*, vol. 2, no. 10, pp. 634–637, 2016.
- [64] S. Newman, *Building microservices: designing fine-grained systems*. " O'Reilly Media, Inc.", 2015.
- [65] R. Dörbecker and T. Böhmman, "The concept and effects of service modularity—a literature review," in *2013 46th Hawaii International Conference on System Sciences*. IEEE, 2013, pp. 1357–1366.
- [66] E. Pluciennik and K. Zgorzałek, "The multi-model databases - a review," in *International Conference: Beyond Databases, Architectures and Structures*. Springer, 2017, pp. 141–152.

- [67] Z. H. Liu, B. Hammerschmidt, D. McMahon, Y. Liu, and H. J. Chang, "Closing the functional and performance gap between sql and nosql," in *Proceedings of the 2016 International Conference on Management of Data*, 2016, pp. 227–238.
- [68] D. Petković, "Json integration in relational database systems," *Int J Comput Appl*, vol. 168, no. 5, pp. 14–19, 2017.
- [69] F. R. Oliveira and L. del Val Cura, "Performance evaluation of nosql multi-model data stores in polyglot persistence applications," in *Proceedings of the 20th International Database Engineering & Applications Symposium*, 2016, pp. 230–235.
- [70] E. Naiburg, E. J. Naiburg, and R. A. Maksimchuck, *UML for database design*. Addison-Wesley Professional, 2001.
- [71] T. J. Teorey, *Database modeling and design: The entity-relationship approach*. Morgan Kaufmann Publishers Inc., 1990.
- [72] M. Kolonko and S. Müllenbach, "Polyglot persistence in conceptual modeling for information analysis," in *2020 10th International Conference on Advanced Computer Information Technologies (ACIT)*. IEEE, 2020, pp. 590–594.
- [73] A. H. Chillón, D. S. Ruiz, and J. G. Molina, "Athena: A database-independent schema definition language," in *International Conference on Conceptual Modeling*. Springer, 2021, pp. 33–42.

Computer Vision: The Effectiveness of Deep Learning for Emotion Detection in Marketing Campaigns

Shaldon Wade Naidoo, Nalindren Naicker*, Sulaiman Saleem Patel and Prinavin Govender
Department of Information Systems,
Durban University of Technology,
Durban, South Africa

Abstract—As businesses move towards more customer-centric business models, marketing functions are becoming increasingly interested in gathering natural, unbiased feedback from customers. This has led to increased interest in computer vision studies into emotion recognition from facial features, for application in marketing contexts. This research study was conducted using the publicly-available Facial Emotion Recognition 2013 data-set, published on Kaggle. This article provides a comparative study of four deep learning algorithms for computer vision application in emotion recognition, namely, Convolution Neural Network (CNN), Multilayer Perceptron (MLP), Recurring Neural Network (RNN), Generative Adversarial Networks (GAN) and Long Short-Term Memory (LSTM) models. Comparisons between these models were done quantitatively using the metrics of accuracy, precision, recall and f1-score; as well and qualitatively by determining goodness-of-fit and learning rate from accuracy and loss curves. The results of the study show that the CNN, GAN and MLP models surpassed the data, and the LSTM model failed to learn at all. Only the RNN adequately learnt from the data. The RNN was found to exhibit a low learning rate, and the computational intensiveness of training the model resulted in a premature termination of the training process. However, the model still achieved a test accuracy of up to 72%, the highest of all models studied, and it is possible that this could be increased through further training. The RNN also had the best F1-score (0.70), precision (0.73) and recall (0.73) of all models studied.

Keywords—computer vision; deep learning; emotion detection; generative adversarial networks, marketing campaigns component

I. INTRODUCTION

The evolution of technology has enabled businesses to gain a better understanding of their customers and develop products and services to better accommodate their markets [1]. The growth of customer centric business models has led to a variety of research studies in business and academia. Notably, Zaim et al in [1] have studied different analytical models for measuring customer experience (CX). A common requirement of these analytical models is that subjective opinions on CX are needed. These opinions are typically captured using surveys and similar methods in [2] and [3]. However, these methods are subject to bias due to human elements such as dishonesty and pressure. This has motivated for more technologically driven, automated approaches to soliciting customer feedback which mitigate some of these biases. One such approach is the use of computer vision to recognise human emotion from facial features. In these studies, cameras were used to record the facial expressions of individuals in real-time. The video

footage was then processed to determine emotions based on the facial expressions that were recorded, and hence obtained a more authentic source of CX feedback.

In this work, we evaluate and compare a variety of machine learning algorithms that perform human emotion recognition in the context of marketing campaigns. Our study focuses on the following three: happiness, anger and surprise. In the context of marketing studies, happiness is an emotion that indicates that adverts are received favourably, which strengthens brand reputation [2]. Anger is a powerful emotion that can evoke responses from communities, and can be leveraged in social marketing campaigns around sensitive or political issues [2]. Finally, the emotion of surprise creates a sense of urgency in a marketing campaign, and may result in consumers taking more immediate action (e.g. online popup advertisements that are valid for a limited time) [2]. Emotion recognition is a classification problem, which literature indicates can be solved using the “deep learning” class of machine learning algorithms [4], [5]. Deep learning is a class of machine learning algorithms that mimics the way the human brain functions, and are typically used in solving prediction or classification problems [6]. The review study by Liang et al. in [7] showed that the convolutional neural network (CNN) deep learning algorithm is popular when performing facial expression recognition (FER). In this paper, other deep learning algorithms were explored and compared to the CNN model to perform FER; viz. the Multilayer Perceptron (MLP), Generative Adversarial Networks (GAN), Long-Short term memory (LSTM) and Recurring Neural Network (RNN) models were studied. The performance of these models was evaluated using accuracy, f1-score, precision, and recall as metrics [8].

May 24, 2022

II. LITERATURE REVIEW

In their 2017 study, Siau and Yang in [6] stated that marketing is a field that has been impacted by advanced technology. The pair elaborated that the marketing field is the most vulnerable to be radicalized by the Fourth Industrial Revolution. In the years following the article, there were significant findings that were made in which the above predictions were fulfilled. Articles that prove this are discussed in the literature review that follows.

Ładyżyński, Żbikowski in [9] experimented in 2019 on a system that directs marketing campaigns to targeted customers,

utilising machine learning techniques. This paper proposed the use of several machine learning algorithms such as Random Forest classifiers, Deep Belief Networks and Classification and regression (CART) classifiers. The team made use of a time series approach. In addition, in order to get results, a marketing campaign simulator was used. This simulator mimicked the call centre scenario described in the paper. Metrics such as precision and recall were applied to evaluate the models.

In the year 2020, Koehn, Lessmann in [10] published findings in which predictions were made about the online shopping behaviour from click-stream data using deep learning. In the paper, the data of user sessions were utilised and mined within a time frame of three months. The data was submitted into various classification models such as Multilayer Perceptron, Long Short-Term Memory (LSTM), Gradient Boosting and Gated Recurrent Networks. The results from this model were based on revenue, the amount of revenue each correct prediction brought in and model accuracy.

Cheng and Tsai in [11] published results in 2019, in which three deep learning models were applied for automated sentiment analysis using social media data. The use of WOM (word of mouth data) was supplied into a LSTM, Bidirectional LSTM (BiLSTM) and Gated Recurrent (GRU) model. The methodology consisted of employing a web crawler in order to gain the data. The data underwent pre-processing due to anomalies such as colloquialism and emotions. The data was labelled using tools such as NTLK and MS text block API, embedded through GLoVe and eventually distributed into the three models. The models were equipped to predict sentiment, and were evaluated using metrics such as precision, recall, F-measure and accuracy [11].

Reviewing these four articles and the respective methodologies proved that Siau and Yang's findings in [11] that was published in their article, which suggests technology is taking over the marketing field at an alarming rate. A common topic among these four articles was the use of Deep Learning. Deep learning is the ability to mimic behaviours of the human brain to solve problems using technology. Emotion depicts a person's feelings about a certain situation and is often portrayed by a facial expression. Therefore, if we were to detect a person's facial expression, one can classify how a person ought to feel about certain scenarios. In the next section of our literature re-view, the relevant findings on the use of deep learning in emotion detection are going to be explored further.

Ko in [12] published a brief review of facial emotion recognition methods. This paper stated that there are two types of ways in which facial recognition can be detected using static and dynamic images. Further on into the paper, the process of recognizing these emotions were described. The process flows from the input images, followed by facial detection and landmark detection, feature extraction and lastly emotion classification were discussed. This process is a standard for all facial emotion recognition experiments; however, it may be flawed. Due to the process the raw image itself is not used and only features are extracted. There are many different models in which can be used to detect facial emotion recognition.

The use of a CNN model for Facial emotion recognition (FER) is common. Tümen, Söylemez in [13] used a CNN model in their paper in which had a 57% success rate. The

methodology included using images from the FER 2013 data-set. Subsequently, the images were supplied into a 3-layer CNN model after feature extraction. The draw-back from this experiment was the data split. Tümen, Söylemez in [13] split the FER data-set into an 80% for training 10% for testing and 10% for validation.

An MLP model is a multi-layer perceptron which consists of multiple neural layers. This model, as used by Tarnowski, Kołodziej in [4] in their findings, is the second most common FER model. Tarnowski, Kołodziej in [4] published findings in which real time 3D data was used to classify amongst the MLP model. The input consisted of a data size of 12 men, in which is relatively small and biased. However, the team was able to obtain a model accuracy of 73%. The methodology used included using a Microsoft Kinect camera in which 12 men sat two metres away from each other. Each participant was tasked with mimicking certain facial expressions. The data collected was deposited into the model, which consisted of seven neural layers, and finally a classification was done. The use of live data left room for many flaws. The response time of the participant to mimic expressions, the lighting conditions, The limitations of the Kinect camera and the size and biases of the data-set were a part of several flaws in this title.

GAN models are an interesting topic. Yi, Sun in [5] experimented on image augmentation using a GAN model. Images were used from the FER 2013 data-set in which a GAN model was able to augment some images and used the images normally. In the paper, it was found that a higher accuracy was generated by using the GAN for image augmentation. However, in this paper, a CNN was done to generate these results as opposed to a GAN model. These models do have potential in facial recognition as stated by Luo, Zhu in [14].

Another deep learning model which is overlooked due to CNNs are LSTM-RNN models. Sepas-Moghaddam, Etemad in [15] stated in their paper that the tested method of an LSTM-RNN model proved to be superior to the normal CNN way of image classification. In the publication, the use of a CNN VGG16 model extracted special features and also added an attention mechanism, which enables effective learning, before running the image through a Bi-directional LSTM-RNN model. The experiment used 800 images from the LFFD data-set. In the absence of the attention layer, the normal LSTM model produced an accuracy of 80%. There is room for improvement with this model. A bigger data-set with more expressions is recommended to solidify the results.

The manner in which technology has impacted the marketing field and how deep learning has been used to identify facial expressions and classify it as an emotion has been expressed in aforementioned statements. Reyes in [16] stated in an article that expression can be derived from emotion, and that emotion is a type of universal language, which stands in the gap for verbal communication. This leads us onto the last section of this literature review. The following section will discuss emotion detection in Marketing.

Yolcu et al in [17] published findings about a deep learning face analysis system to monitor customer interest. In their findings they monitored customers head position and facial expression to determine how interested customers are. In the methodology of the article Yolcu et al, used the Viola and

Jones algorithm for feature extraction, in which the image was cropped and only the head pose, and facial expression was extracted. A 3-layer CNN model was used to classify 8040 images for head position and 1206 for facial expression, from the Radboud Face Database. To verify this system the KDEF database was used with 4900 images and seven different emotions. A 90:10 split was done for model training and testing. A 94% accuracy was achieved for emotion accuracy in this article. Yolcu et al, used a mask generation flow in order to achieve such high results. The images were also pre-processed twice, using the Viola and Jones algorithm and also another CNN model. To improve this research, a multimodal approach in which multiple models can be used are suggested.

Ceccacci, Generosi in [3] experimented on a deep learning system which tracked and monitored customer behaviour in store. The system conducted testing in real life, using a Logitech Quickcam 4K camera. This experiment input the images of customers into these cameras which was linked to a CNN model. A total of 30 customers, split equally of gender, were tested. Images of them were distributed into a CNN model that was trained and tested using FER+ and EmotioNet data-sets in which consists of millions of images between the two. Model accuracy was not stated in the article. However, the model was able to predict 66% of emotions in the live experiment. These are two of the articles in which face emotion recognition was processed and completed using deep learning.

In our review of similar work, we have discussed deep learning in marketing, the use of deep learning in emotion detection and emotion detection in marketing. It can be assumed that there are gaps in literature that can add to how one can utilise deep learning in emotion recognition in order to aid marketing strategies.

III. METHODS AND MATERIALS

A. Data Acquisition

This study uses the Facial Emotion Recognition 2013 (FER 2013) data-set, available publicly on Kaggle [16]. The FER 2013 data-set consists of 35 685 greyscale images of size 48x48 pixels, and the provides recommended splits for testing and training image sets.



Fig. 1. Sample Images: Angry

In our study, the original FER 2013 data-set was first divided into two smaller data-sets. This was a result of limited processing power available by the HP i5 4GB RAM that was

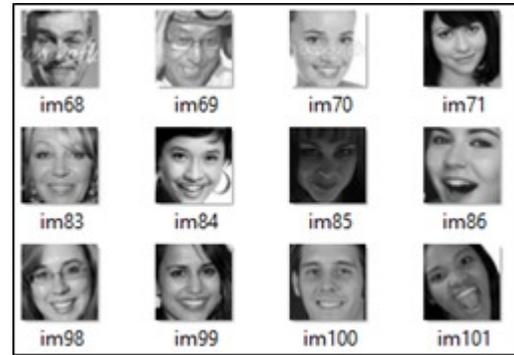


Fig. 2. Sample Images: Happy

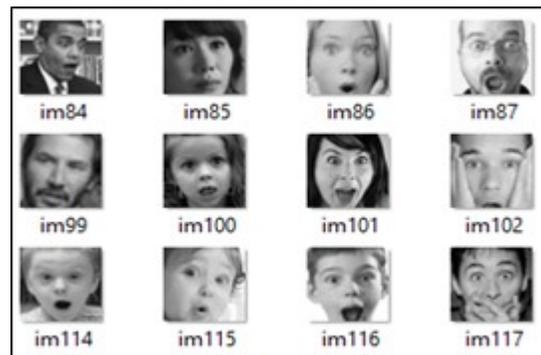


Fig. 3. Sample Images: Surprised

used in performing these experiments. Both data-sets were then subdivided into training and testing sets, using an 80:20 ratio. In this research experiment, three emotions from the original FER 2013 data-set were considered (i.e. happy, angry, and surprised). Sample images from the FER 2013 data-set are provided in Fig. 1, 2 and 3, which show sample images of angry, happy and surprised facial expressions, respectively. The motivation for selecting these three emotions for the scope of this study was given in Section I.

The first data-set contained 12 000 images from these three images classes, consisting of 3 000 training images and 1 000 test images per emotion class. A similar structure was used for the second data-set, in order to avoid any type of bias. This data-set consisted of 21 000 images (5 600 training images and 1 400 testing images per emotion class).

B. Data Pre-Processing

There are generally two approaches to performing facial expression recognition (FER) using computer vision. The first is to pre-process images and extract features from them, which form the input to the machine learning algorithm. This has been shown to provide robust machine learning models in some studies, for example the study by Ceccacci et al [3]. The second approach is to use raw pixel data directly, which allow deep learning models to identify features for themselves during the model training process [18]. Exploratory checks performed on the data indicated that quality of images was high, and hence the experiment was designed using raw pixel data with no feature extraction or cleaning. However, model-

TABLE I. EXPERIMENT PROCESS

Step No.	Step Name	Details
1	Imports	Relevant import classes were imported into the experiment. Consistent imports namely would be Tensorflow, Keras, Matplotlib.pyplot, OpenCV and OS.
2	Pre-processing	A method is created in which sorts data according to its label and expression and populates them into an array. Images are also resized in this method
3	Train and Test variable population	The above method created populates the relevant data into their train and test variables.
4	Array Reshape	The array size of each image is reshaped according to the model requirements.
5	Hyper-parameters set	A model is called in which hyper parameters are set. All models consist of the same hyper-parameters, except for layers and epochs. This is due to model requirements.
6	Model compiled	The model is compiled and set to produce model accuracy. Categorical cross-entropy to help us measure the performance of the classification output and tell us the difference between the classifications.
7	Epochs set, model fitted.	Epoch numbers are set and the model is fitted with training and testing data. The output of this step would be the accuracy of the model.
8	Graphs and Table of results.	Graphs are created to measure validation loss and accuracy. Table of results calculated

specific data transformations were necessary (discussed further in subsection 4).

According to [19], data segmentation and data transformation are critical factors in data pre-processing. Consequently, in this experiment, data segmentation was performed first by splitting the training and testing data to ensure we kept the integrity of the 80:20 split. Furthermore, each emotion had been allocated its own files within the testing and training folders. Each image was resized using OpenCV, to ensure the data would meet model requirements. In order to ensure the deep learning model was able to learn properly, each image was converted into an array. In which, the label and feature of each image was stored respectively.

Singh and Singh in [20] stated in their article that normalization of data in deep learning is imperative in order to make a good contribution to each feature. The writers also stated that normalization is a critical success factor in the learning of each algorithm. Therefore, in our experiment, we normalize the data according to the requirement of the algorithm being used.

Concluding our data pre-processing, the use of methods such as data segmentation, data transformation and data normalization ensured that each model that was used has an equal chance at performing its best.

C. Design of Study

The experimental design method was used in the study of the effectiveness of deep learning for emotion detection in marketing campaigns. This type of study proved effective in a number of deep learning articles [4], [21], [22]. The independent variable being the different types of deep learning models such as CNN, GAN, MLP, RNN and LSM. The dependent variable was our metrics such as model accuracy, f1-score, precision and recall. The model accuracies depend on the type of model in order to prove the title of our paper. Experiments carried out on both data-sets followed the same steps and procedures, to prevent biasness in any way. Our experiments were conducted on Jupyter Notebook, using Python 3.0. It consisted of eight steps, as indicated in Table I above.

The steps shown Table I were a general guideline of how we carried out the experiment on how effective deep learning

was used for emotion recognition. The process outlined is consistent with the extant literature, such as in [23].

D. Algorithms

In the study that was conducted, we focused on five deep learning models. Convolutional Neural Network (CNN), Long Short-Term Memory (LSTM), Multilayer Perceptron (MLP), Generative Adversarial Network (GAN) and Recurrent Neural Networks (RNN). In this subset of our paper, we will be discussing in depth, these models and their unique capabilities they bring to the deep learning field. We will also discuss how they have been implemented and define functions used to generate appropriate outputs.

1) *Convolutional Neural Network (CNN)*: CNN classifiers have been getting high amount of recognition in the deep learning world [24]. Most image classification problems use CNN models, due to the high levels of model accuracy, CNN has also proved to be the better and most preferred model amongst most deep learning models for image classification [25]. Convolutional neural networks generally consist of two layers. The first being a convolution layer, also known as a C layer. The second layer being the subsampling layer, known as the S layer. Each S layer follows a C layer as depicted in graph below. Convolutional Neural Networks have an advantage over normal deep learning models, having the ability to accept 2D images without major changes to the array.

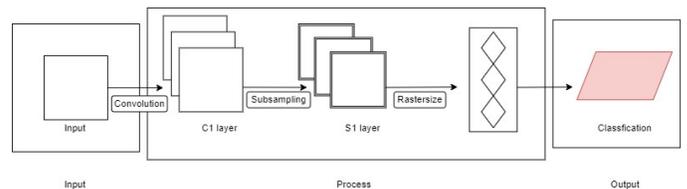


Fig. 4. Conventional CNN

The illustration depicted in Fig. 4 is a basic CNN model. Normally CNN models consist of two C1 layers and two S1 layers. However, to explain its structure we will use a single C1 and S1 layer. The input image, usually in the form of an array, would be fed into the first C1 layer. In this layer, feature maps would be formed through the convolutions. These feature maps would be inputted next into the S1 layer. During sampling the

feature map size would be reduced by a pooling method which normally consists of 2×2 . Usually this process is repeated, depending on the number of layers. After the S1 layer the data is rasterized and a classification is formed.

The CNN model that we experimented on consisted of five convolution layers, in which had a kernel size of 3. Five Subsampling layers of pooling size 2×2 . We used the Rectified Linear Unit (ReLU) activation function to output the input only if the output is a positive value. We used batch normalization to improve the speed of our model and finally a dropout of 0.20 is used in order to prevent overfitting. The equation for the ReLU activation function is describe as:

$$f(x) = \max(0, x) \quad (1)$$

2) *Multi-Layer Perceptron (MLP)*: MLP models are binary classifiers in the field of deep learning. General MLP models consist of multilayers however, single layers are also used. MLP models are famously used to state whether an input is something or not. Each Perceptron model consists of three layers. An input layer, a hidden and an output layer. Perceptron has a general rule, it states that the model will learn the best weight coefficients.

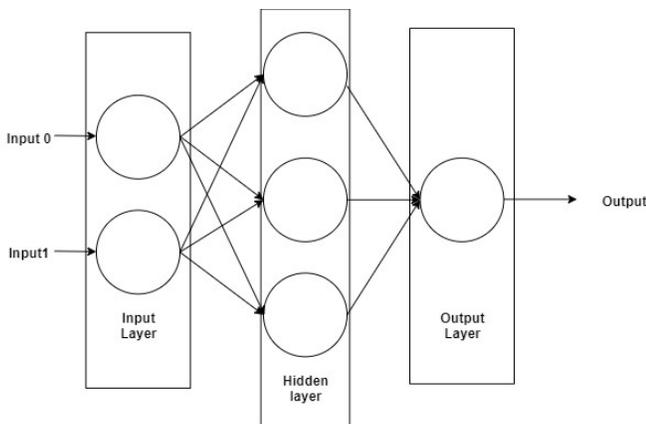


Fig. 5. MLP Model

Following from the beginning of the diagram in Fig. 5 the inputs were read in the input layer. In the hidden layer the weights were then calculated and their net input function was determined. After this, their activation function transformed the net input into an output, within the output layer. If any errors occurred, they were caught and sent back to the input layer. The perception activation function is described as:

$$f(x) = \begin{cases} 1 & \text{if } \mathbf{wX} + \mathbf{b} > 0 \\ 0 & \text{otherwise} \end{cases}, \quad (2)$$

where \mathbf{w} is the real weight value matrix, \mathbf{b} is the bias vector and $\mathbf{X} = [x_1 \dots x_k]^T$ is the k -dimensional vector of input data. Note that the boldface in (2) represents matrix variables, rather than scalar variables and that $(\cdot)^T$ represents the matrix transpose.

As stated above, each MLP model consists of three layers. In the research that we have carried out, the MLP model that we used consists of two dense layers. The first dense layer

having 128 neurons and using the ReLU activation function. The second layer consisting of 5 neurons and the softmax activation function

E. Generative Adversarial Network (GAN)

The popularity around GAN models stems around their ability to augment data. GAN models are used famously throughout the deep learning fields to enhance other deep learning algorithms [26]. GANs are generally made up of two smaller deep neural networks. The first being a generator, in which is responsible for generating data. The second being a discriminator, which takes the real and fake images to classify which one is real or fake.

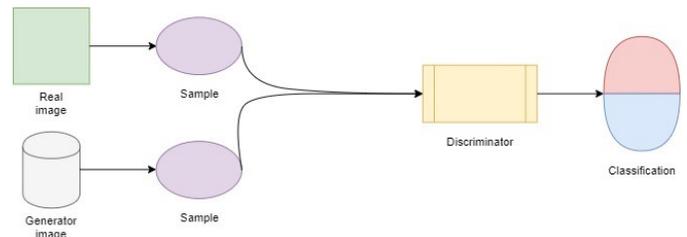


Fig. 6. GAN Model

Fig. 6 depicts the general structure of a GAN model. In the experiment that we have conducted, the GAN model does not consist of a generator, due to the input images already being available. However, the discriminator that we have used consists of a method created. The discriminator consist of two down sample layers consisting of three kernels each and LeakyReLU as an activation function, which was suggested from reviewing the works of Krestinskaya, Choubey in [27]. The activation function for LeakyReLU is described below.

$$f(x) = \max(0.01x, x), \quad (3)$$

where x is the input to the activation function.

Our discriminator also has a flatten feature, followed by a 0.4 dropout layer. The output layer of our GAN model has a softmax activation function. It is understood that GAN models classify whether an image is fake or not. However, in our research we were able to recognize whether an emotion was present or not by using three GAN models developed independently. In this study we utilized the GAN models to independently test for the emotions “happy”, “angry” and “surprised”. The structure of the GAN was based on the work in [28].

F. Recurring Neural Network (RNN)

The operations and structures of RNNs are the same as Fig. 7. RNNs feed the information through their vectors back into their input gates. In the RNN model that we used, the hidden layer consisted of 64 neurons, a dropout layer and a recurrent dropout layer. In the second hidden layer we used the ReLU activation function. For the output layer we use a dropout of 0.5 and the softmax activation function.

G. Long Short-Term Memory (LSTM)

LSTM models are a type of RNN deep learning model. Conventional neural networks have a feed forward tendency. However, the LSTM models have a feedback connection. LSTM neural networks are advancements of the general Recurrent neural network model and are known for modelling chronological sequences [29]. However, they can be used to classify images [30].

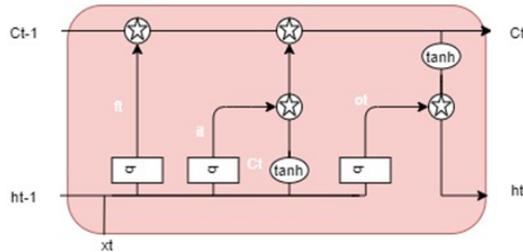


Fig. 7. General LSTM Model

The labels C_{t-1} , h_{t-1} and x_t represent the input values for the LSTM diagram shown in Fig. 7. The stars represent pointwise operations, and the rectangular boxes are neural networks. Each arrow represents a vector movement. C_{t-1} to C_t represents the cell state and is the key to any LSTM model. Each pointwise operation is followed by a quadrature which were seen by the sigmoid function. Together the pair is called a gate, and their prime purpose is to allow data to flow through. Each LSTM model consisted of three gates to control the flow of information. The outputs of these gates were either 0 or 1. Which represent to let information through or not, respectively. Information flowed from h_{t-1} and x_t through the gates one and two, followed by the hyperbolic tangent operator. Then finally through the third gate and the second hyperbolic tangent operator followed by the output [31].

Our LSTM model has two layers, both of which uses the ReLU activation function, and a kernel optimizer. The model also has a loss and momentum function to prevent over fitting and help with model performance.

H. Performance Metrics

In the experiment that was conducted a total of four metrics were used. The descriptions of the evaluation metrics are [8]:

- **Precision:** Precision is the number of most relevant instances amongst those that were retrieved. It is a metric used in our findings to determine how correct our model is and how many true predictions were made.
- **Recall:** Recall measures the actual number of relevant instances made, without taking those that were retrieved into accord [8]. This metric determines the number of positive predictions made by the model.
- **F1-score:** Also known as the F score, these metrics are used in binary classifiers. They determine how accurate each model is according to the data-set.
- **Model accuracy:** This is the output accuracy taken from the model training. It determines how accurately the model has been trained.

IV. RESULTS AND DISCUSSION

In this section, we present the results of the comparative study that investigated emotion recognition using deep learning algorithms. Five deep learning models were considered (i.e. CNN, MLP, GAN, RNN and LSTM) and compared in terms of four metrics (accuracy, F1-score, precision and recall). The learning rates of each model produced was also discussed. Details on the configuration of each algorithm studied was provided in Section III-D.

The results of the study are summarised in Table II, which gives results broken down by model and data set used. Analysing the results presented in Table II, the following points are noted:

- The LSTM model fails to learn for both data-sets and is not suitable for performing FER. This is intuitively understandable, as the LSTM model is deep learning algorithm that is designed primarily for sequential or chronological data [29]. As such, poor performance was understandable when considering independent, unrelated images such as those used in this study.
- The CNN, MLP and GAN models are overfitted to the respective data-sets and do not produce trained models that can be applied to generalised data. This makes these models unsuitable for performing facial emotion recognition.
- The RNN learns appropriately and achieves a 72% accuracy on the larger of the two data-sets (Data-set 2). The accuracy and loss curves indicate that this accuracy can be improved by further training, but this was not done in this study due to the computational intensiveness of training the model and its low learning rate. The RNN performs best in terms of precision, recall and F1-score.

V. CONCLUSION

In this study, we considered five deep learning algorithms for facial emotion recognition, with the overall objective of utilising deep learning to improve marketing business functions by soliciting more accurate feedback on CX. The algorithms studied were the CNN, MLP, GAN, RNN and LSTM. Although literature often uses the CNN for facial emotion recognition studies, in our study the CNN overfits to training data and the RNN is found to be more suitable. The designed RNN is computationally intensive to train, and training in this study was terminated prematurely. The model achieved a 72% testing accuracy on a 21 000-image subset of the FER 2013 data-set, and indicators are that more training could improve this accuracy. This motivates for more intense studies that design RNN-based computer vision systems for facial emotion recognition. We recommend that future works in this area use cloud computing technologies to overcome the limitations of computational intensiveness. In so doing, a larger and more varied data-set could be evaluated which may influence the performance of the models evaluated. It would also be interesting to consider more emotion classes in future studies to cover a broader spectrum of human reactions.

TABLE II. SUMMARY OF RESULTS

Model	Data-set	Accuracy	Precision	F1-score	Recall	Learning and Fit
CNN	1	0.68	0.73	0.68	0.68	Loss-epoch curve shows evidence of severe overfitting.
	2	0.70	0.70	0.70	0.70	Loss-epoch curve shows evidence of severe overfitting.
MLP	1	0.56	0.66	0.57	0.57	Loss-epoch curve shows evidence of moderate overfitting.
	2	0.57	0.63	0.56	0.56	Loss-epoch curve shows evidence of severe overfitting.
GAN	1	0.68	0.67	0.60	0.60	Loss-epoch curve shows evidence of moderate overfitting.
	2	0.65	0.72	0.66	0.66	Loss-epoch curve shows evidence of severe overfitting.
RNN	1	0.60	0.78	0.71	0.71	Learning rate is slow, further training could improve accuracy.
	2	0.72	0.70	0.73	0.73	Learning rate is slow, further training could improve accuracy.
LSTM	1	0.33	0.50	0.33	0.33	Model does not learn.
	2	0.35	0.30	0.46	0.46	Model does not learn.

FUNDING

All funding to support this research was provided by the Durban University of Technology.

DATA AVAILABILITY STATEMENT

The data on facial emotion recognition 2013 (FER-2013) is available online at <https://www.kaggle.com/ananthu017/emotion-detection-fer>

ACKNOWLEDGMENT

The authors would like to acknowledge Lerissa Gounden (BEd) of the University of Johannesburg for editing and Ghulam Masudh Mohammed (BICTH) of the Durban University of Technology for their respective inputs.

CONFLICTS OF INTEREST

The authors declare that there are no conflicts of interest.

REFERENCES

- [1] H. Zaim, M. Ramdani, and A. Haddi, "E-crm success factors as determinants of customer satisfaction rate in retail website," *International Journal of Computer Information Systems and Industrial Management Application*, vol. 12, pp. 082–092, 2020.
- [2] S. Harvey, "The power of emotional marketing: Once more with feeling," <https://fabrikbrands.com/the-power-of-emotional-marketing/>, n.d., accessed: 2022-04-02.
- [3] S. Ceccacci, A. Generosi, L. Giraldi, and M. Mengoni, "An emotion recognition system for monitoring shopping experience," in *Proceedings of the 11th PErvasive Technologies Related to Assistive Environments Conference*, 2018, pp. 102–103.
- [4] P. Tarnowski, M. Kołodziej, A. Majkowski, and R. J. Rak, "Emotion recognition using facial expressions," *Procedia Computer Science*, vol. 108, pp. 1175–1184, 2017.
- [5] W. Yi, Y. Sun, and S. He, "Data augmentation using conditional gans for facial emotion recognition," in *2018 Progress in Electromagnetics Research Symposium (PIERS-Toyama)*. IEEE, 2018, pp. 710–714.
- [6] K. Siau and Y. Yang, "Impact of artificial intelligence, robotics, and machine learning on sales and marketing," in *Twelve Annual Midwest Association for Information Systems Conference (MWAIS 2017)*, vol. 48, 2017, pp. 18–19.
- [7] H. Liang, X. Sun, Y. Sun, and Y. Gao, "Text feature extraction based on deep learning: a review," *EURASIP journal on wireless communications and networking*, vol. 2017, no. 1, pp. 1–12, 2017.
- [8] G. M. Mohamed, S. S. Patel, and N. Naicker, "Data Augmentation for Deep Learning Algorithms that perform Driver Drowsiness Detection," *Scientific Programming*, 2022, under review.
- [9] P. Ładyżyński, K. Żbikowski, and P. Gawrysiak, "Direct marketing campaigns in retail banking with the use of deep learning and random forests," *Expert Systems with Applications*, vol. 134, pp. 28–35, 2019.
- [10] D. Koehn, S. Lessmann, and M. Schaal, "Predicting online shopping behaviour from clickstream data using deep learning," *Expert Systems with Applications*, vol. 150, p. 113342, 2020.
- [11] L.-C. Cheng and S.-L. Tsai, "Deep learning for automated sentiment analysis of social media," in *Proceedings of the 2019 IEEE/ACM international conference on advances in social networks analysis and mining*, 2019, pp. 1001–1004.
- [12] B. C. Ko, "A brief review of facial emotion recognition based on visual information," *sensors*, vol. 18, no. 2, p. 401, 2018.
- [13] V. Tümen, Ö. F. Söylemez, and B. Ergen, "Facial emotion recognition on a dataset using convolutional neural network," in *2017 International Artificial Intelligence and Data Processing Symposium (IDAP)*. IEEE, 2017, pp. 1–5.
- [14] Y. Luo, L.-Z. Zhu, and B.-L. Lu, "A gan-based data augmentation method for multimodal emotion recognition," in *International Symposium on Neural Networks*. Springer, 2019, pp. 141–150.
- [15] A. Sepas-Moghaddam, A. Etemad, F. Pereira, and P. L. Correia, "Facial emotion recognition using light field images with deep attention-based bidirectional lstm," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 3367–3371.
- [16] J.-F. Reyes, "Effect of emotion on marketing landing page conversion," Ph.D. dissertation, University of Baltimore, 2016.
- [17] G. Yolcu, I. Oztel, S. Kazan, C. Oz, and F. Bunyak, "Deep learning-based face analysis system for monitoring customer interest," *Journal of ambient intelligence and humanized computing*, vol. 11, no. 1, pp. 237–248, 2020.
- [18] H. Zhang, A. Jolfaei, and M. Alazab, "A face emotion recognition method using convolutional neural network and image edge computing," *IEEE Access*, vol. 7, pp. 159 081–159 089, 2019.
- [19] X. Zheng, M. Wang, and J. Ordieres-Meré, "Comparison of data preprocessing approaches for applying deep learning to human activity recognition in the context of industry 4.0," *Sensors*, vol. 18, no. 7, p. 2146, 2018.
- [20] D. Singh and B. Singh, "Investigating the impact of data normalization on classification performance," *Applied Soft Computing*, vol. 97, p. 105524, 2020.
- [21] M. M. T. Zadeh, M. Imani, and B. Majidi, "Fast facial emotion recognition using convolutional neural networks and gabor filters," in *2019 5th Conference on Knowledge Based Engineering and Innovation (KBEI)*. IEEE, 2019, pp. 577–581.
- [22] N. Jain, S. Kumar, A. Kumar, P. Shamsolmoali, and M. Zareapoor, "Hybrid deep neural networks for face emotion recognition," *Pattern Recognition Letters*, vol. 115, pp. 101–106, 2018.
- [23] P. Singh, "Visual-Recognition-using-CNN," <https://github.com/pradeepsingh/Visual-Recognition-using-CNN/>, 2018, accessed: 2022-04-09.
- [24] D. Han, Q. Liu, and W. Fan, "A new image classification method using cnn transfer learning and web data augmentation," *Expert Systems with Applications*, vol. 95, pp. 43–56, 2018.
- [25] S. S. Yadav and S. M. Jadhav, "Deep convolutional neural network based medical image classification for disease diagnosis," *Journal of Big Data*, vol. 6, no. 1, pp. 1–18, 2019.
- [26] Z. Peng, J. Li, and Z. Sun, "Emotion recognition using generative adversarial networks," in *2020 International Conference on Computer Engineering and Intelligent Control (ICCEIC)*. IEEE, 2020, pp. 77–80.

- [27] O. Krestinskaya, B. Choubey, and A. James, "Memristive gan in analog," *Scientific reports*, vol. 10, no. 1, pp. 1–14, 2020.
- [28] S. Shamsheer, "Digit-recognition-using-GAN," <https://github.com/SaraShamsheer/Digit-recognition-using-GAN>, 2020, accessed: 2022-04-09.
- [29] N. K. Manaswi, N. K. Manaswi, and S. John, *Deep learning with applications using python*. Springer, 2018.
- [30] J. H. Bappy, J. R. Barr, N. Srinivasan, and A. K. Roy-Chowdhury, "Real estate image classification," in *2017 ieee winter conference on applications of computer vision (wacv)*. IEEE, 2017, pp. 373–381.
- [31] A. Sinha, "Understanding of LSTM Networks," <https://www.geeksforgeeks.org/understanding-of-lstm-networks/>, 2021, accessed: 2022-04-02.

Using Machine Learning Techniques to Predict Bugs in Classes: An Empirical Study

Musaad Alzahrani
Department of Computer Science
Albaha University
Al-Baha 65799, Saudi Arabia

Abstract—Software bug prediction is an important step in the software development life cycle that aims to identify bug-prone software modules. Identification of such modules can reduce the overall cost and effort of the software testing phase. Many approaches have been introduced in the literature that have investigated the performance of machine learning techniques when used in software bug prediction activities. However, in most of these approaches, the empirical investigations were conducted using bug datasets that are small or have erroneous data leading to results with limited generality. Therefore, this study empirically investigates the performance of 8 commonly used machine learning techniques based on the Unified Bug Dataset which is a large and clean bug dataset that was published recently. A set of experiments are conducted to construct bug prediction models using the considered machine learning techniques. Each constructed model is evaluated using three performance metrics: accuracy, area under the curve, and F-measure. The results of the experiments show that logistic regression has better performance for bug prediction compared to other considered techniques.

Keywords—Software bugs; bug prediction; machine learning techniques; software metrics; unified bug dataset

I. INTRODUCTION

Software development is an error-prone process. Mistakes and errors that occur during the development process result in bugs that can ultimately cause software failure [1]. Software testing is one of the most important phases in the software development process which aims to identify bugs and to ensure the overall quality of systems before they are released. However, the testing cost and effort can grow dramatically when the size and complexity of a system increase. It is estimated that the cost of the testing activities constitutes around 25% of the total cost of the software development budget and it can reach to 50% when the size and complexity of the system increase [2], [3]. Therefore, the testing resources should be allocated efficiently in order to minimize the total cost of the overall development process.

Software Bug Prediction (SBP) is one of the most useful techniques that can be used to decrease the testing cost and effort [4]. The main goal of SBP is to identify the modules that are likely to have bugs. Software professionals use SBP models at the beginning of the testing phase to classify the modules of a system into bug-prone modules and non-bug-prone modules based on a set features extracted from the modules. The most commonly used features are software metrics that measures different characteristics of the module such complexity, size, coupling, and cohesion. Fig. 1 shows the basic architecture of a SBP model. Most of the testing cost and effort should be

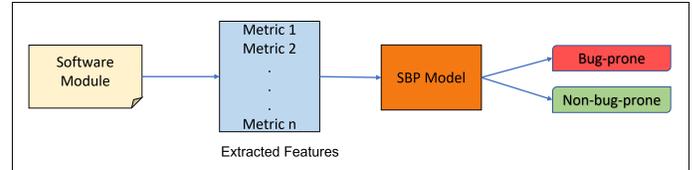


Fig. 1. Basic Architecture of a SBP Model.

allocated on bug-prone modules in order to use the limited testing resources efficiently.

Many approaches have been introduced into the literature that have shown the success of using machine learning algorithms to construct SBP models (e.g., [5], [6], [7], [8]) based on public available bug datasets such as NASA and Columbus [9]. However, the size and quality of the bug datasets used to train SBP models can have a great impact on the performance of the models and can limit their generality. Ferenc, Rudolf et al. [10] explained several issues in most commonly used bug datasets including the missing of the source code elements associated with data; dissimilarities in terms of granularity, features, and format between the datasets; the missing values in the datasets; and the existence of contrasting bug information. To mitigate these issues, they produced a unified bug dataset at class and file level by analyzing the source code of the reported systems in 5 public bug datasets. The Unified Bug Dataset [10] contains the values of 60 metrics and bug information for 47,618 classes and for 43,744 files and their corresponding source code. Although the Unified Bug Dataset is considered to be a large and clean dataset and has better quality compared to most commonly used bug datasets, it has been used in only a few studies in literature (e.g., [10], [11]) to build machine learning based SBP models.

Therefore, this paper empirically investigates the performance of 8 well-known machine learning techniques, namely, Logistic Regression (LR), Support Vector Machine (SVM), K-Nearest Neighbor (KNN), Naive Bayes(NB), Decision Tree (DT), Bagging, Random Forest (RF), and AdaBoost based on the Unified Bug Dataset. The contribution of the paper is twofold:

- Constructing 8 SBP models based on the Unified Bug Dataset.
- Evaluating and comparing the performance of the constructed SBP models.

The rest of the paper is organized as follows. Related

studies are summarized and discussed in Section II. Section III describes the used methodology. The results and discussion are given in Section IV. The threats to validity are presented in Section V. Section VI concludes the study and provides future work.

II. RELATED WORK

SBP refers to the process of predicting buggy modules in a software system. In the last two decades, many approaches in the literature have introduced SBP models that tried to identify a causal relationship between a set of characteristics or features of a given software module and the existence of bugs in that module based on historical bug datasets. The most commonly used features to predict bugs are software metrics such as McCabe's Cyclomatic Complexity [12], Halstead Metrics [13], and Chidamber and Kemerer object-oriented metrics suite [14].

The majority of SBP models are constructed using machine learning algorithms such as LR (e.g., [5], [6]), SVM (e.g., [7], [8]), and KNN (e.g., [8]). The study by Basili et al. [5] was one of the early studies that used LR to construct a SBP model for the purpose of empirically validating Chidamber and Kemerer object-oriented metrics as quality indicators based data collected from 8 small object-oriented systems. The results of the study showed that the object-oriented metrics are beneficial to some degree to predict bugs in software systems in early phases of their development life cycle. Osman et al. [8] studied the impact of adjusting the parameters of two machine learning algorithms: SVM and KNN to build better SBP models. They conducted a set of experiments on five systems and the results showed that tuning the parameters of the two algorithms can improve their accuracy when compared to the default parameters setting. Researchers in [15] experimentally compared the performance of NB to DT for defect prediction based NASA Datasets [9]. Their results suggested that NB is more accurate and useful in predicting software defects when compared to DT. Singh et al. [16] analyzed the performance of five machine learning algorithms namely Artificial Neural Network, Particle Swarm Optimization, DT, NB, and SVM. They carried out a set of experiments on NASA datasets to compare the accuracy of the considered algorithms when used for software defect prediction. The output of the experiments showed that SVM outperformed the other 4 algorithms. Matloob et al. [17] conducted a systematic literature review on software defect prediction using ensemble learning methods. They considered only studies that were published in the period from 2012 to 2021. The results of their systematic literature review showed that RF, boosting, and bagging are the most frequently proposed methods in the literature during the considered period. In addition, their results showed that most of the proposed ensemble models were built based on PROMISE datasets [9] (which consists of a set of public datasets mostly NASA datasets). In a recently published study [11], the performance of the two ensemble learning methods: AdaBoost and Bagging was investigated. A set of experiments were conducted on the Unified Bug Dataset [10]. The results indicated that AdaBoost with a DT as a base learner outperformed Bagging technique.

A considerable amount of previously published studies have indicated the effectiveness of using machine learning algorithms to build SBP models. However, most of these

studies used bug datasets that are small or have erroneous data such as NASA datasets [18], [11]. Therefore, this study tries to bridge this gap by using a set of well-known machine learning algorithms to construct SBP models based on the Unified Bug Dataset [10], which is a large and clean dataset and which has been used in only a few studies in the literature [11].

III. METHODOLOGY

A. Motivation

Many machine learning algorithms have been used in previous studies to construct SBP models. However, the results of these studies are not always agreeing on the superiority of a machine learning algorithm or technique over others. In addition, most of the previous studies built SBP models based on bug datasets (such as NASA datasets) that have been shown to be noisy and containing erroneous data [19], [18], [11], which can have a significant impact on the performance of these models.

Motivated by the previously mentioned remarks, this study aims to answer the following research question:

- RQ: What is the most effective machine learning technique (in terms of performance metrics) for bug prediction in classes based on the Unified Bug Dataset?

B. Research Framework

The overall research framework of this study is depicted in Fig. 2. The input dataset is first preprocessed. The preprocessing step includes data normalization and feature selection. After the preprocessing of the input dataset, the ten-fold cross-validation is used to train and evaluate the considered machine learning models. The ten-fold cross-validation randomly divides the dataset into 10 equal size subdatasets. For each subdataset, the remaining 9 subdatasets are used to train a model and the subdataset is used to test the performance of the model. Finally, the results of the ten-fold cross-validation for each model are averaged and reported.

C. Dataset

The Unified Bug Dataset [10] is used to construct the SBP models. The dataset contains information for 47,618 classes and for 43,744 files and their corresponding source code. The information includes the values of 60 software metrics (including McCabe's Cyclomatic Complexity [12], Halstead Metrics [13], and Chidamber and Kemerer object-oriented metrics suite [14]) and the number of bugs in each class and file. The values of the software metrics for the classes and files were calculated from their source code using the open-source OpenStaticAnalyzer tool [20]. The bug information of the classes and file were collected from 5 public bug datasets namely: PROMISE [21], Eclipse Bug Dataset [22], Bug Prediction Dataset [23], Bugcatchers Bug Dataset [24], and GitHub Bug Dataset [25].

D. Dependent and Independent Variables

The dependent variable Y in this study is binary (i.e., $Y \in \{0, 1\}$) where 0 means that the class does not have a bug (referred as non-buggy class) and 1 means the class has

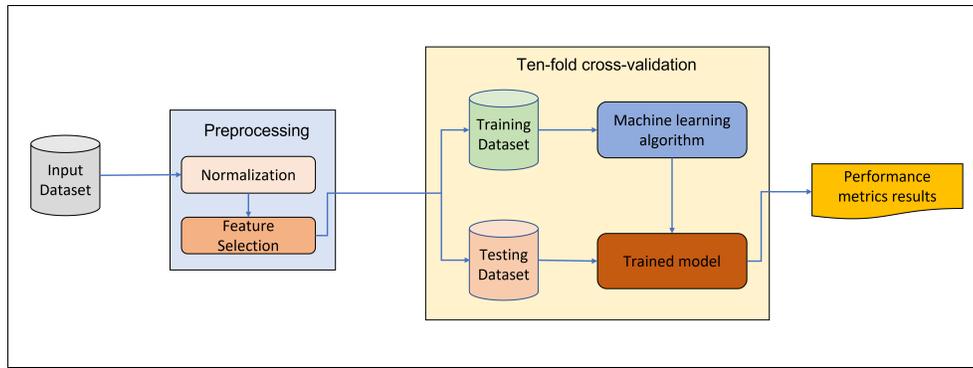


Fig. 2. The Research Framework of this Study.

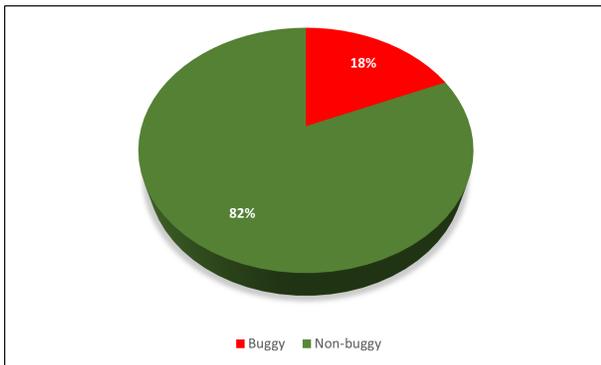


Fig. 3. The Distribution of Classes in the used Dataset.

at least one bug (referred as buggy class). Figure 3 shows the distribution of buggy and non-buggy classes in the Unified Bug Dataset. It can be seen from Fig. 3 that the Unified Bug Dataset has imbalanced distribution in terms of buggy and non-buggy classes. The non-buggy classes make up to 82% of the total classes in dataset whereas the buggy classes form only 18%. The ultimate goal of a SBP model is to label a class in question with 0 or 1 based on the values of a set of independent variables. The software metrics of the classes reported in the Unified Bug Dataset are used as independent variables in this study. Software metrics of a class are quantitative measurements that indicate the degree to which a class possesses a property or attributes such as class complexity, size, coupling and cohesion.

E. Data Preprocessing

Two common techniques are applied sequentially to preprocess the considered dataset namely: MinMaxScaler [26], and correlation-based filter-subset feature selection with BestFirst search [27].

One issue in the Unified Bug Dataset is that the dataset includes software metrics that are not normalized (i.e., they do not have an upper bound) and they differ in the order of magnitude [4]. This issue can have a negative impact on the accuracy of a prediction model [4]. Normalization is a commonly used technique that is used to address this issue by rescaling of the original values of variables to a specific range. In this paper, the MinMaxScaler [26] technique is applied

in the data preprocessing to transform the original values of all the metrics in the Unified Bug Dataset between the closed interval 0 and 1.

High dimensionality is another issue in the Unified Bug Dataset. The dataset includes 60 software metrics. Fig. 4 shows the Spearman correlation coefficients between each pair of software metrics in the dataset. As it can be seen from Fig. 4, some of these metrics have strong correlations with each other. Building SBP models based on high dimensional redundant dataset takes more time and computational resources and can negatively affect the performance models [4], [28]. Therefore, researchers often apply feature selection techniques to address this problem before constructing prediction models [1], [29], [30], [31]. In this study, the correlation-based filter-subset feature selection with best first search is used. This technique was found to be the best when used in the field of SBP among 30 feature selection techniques that were analyzed in a large-scale study [27].

F. Learners

The learners that are used to build SBP models in this study include: Logistic Regression (LR), Support Vector Machine (SVM), K-Nearest Neighbor (KNN), Naive Bayes (NB), Decision Tree (DT), Bagging, Random Forest (RF), and AdaBoost. The former 5 learners referred in the literature as traditional learners whereas the latter 3 learners referred as ensemble learners. In traditional learning techniques, a single learner (e.g., LR) is used to build a prediction model. On the other hand, a combination of learners are used to build a prediction model in ensemble learning techniques [17]. A brief description of each learner is given in the following.

Logistic Regression (LR): LR is a statistical model used to predict a binary dependent variable based on a set of independent variables using the following equation:

$$\pi(X_1, X_2, \dots, X_n) = \frac{1}{1 + e^{-(C_0 + C_1 X_1 + C_2 X_2 + \dots + C_n X_n)}} \quad (1)$$

where X_1, X_2, \dots, X_n are the independent variables and C_1, C_2, \dots, C_n are estimated regression coefficients. The larger the absolute value of the coefficient, the stronger the impact of the independent variable is on the dependent variable. π is the probability that a the dependent variable is 0 or 1.

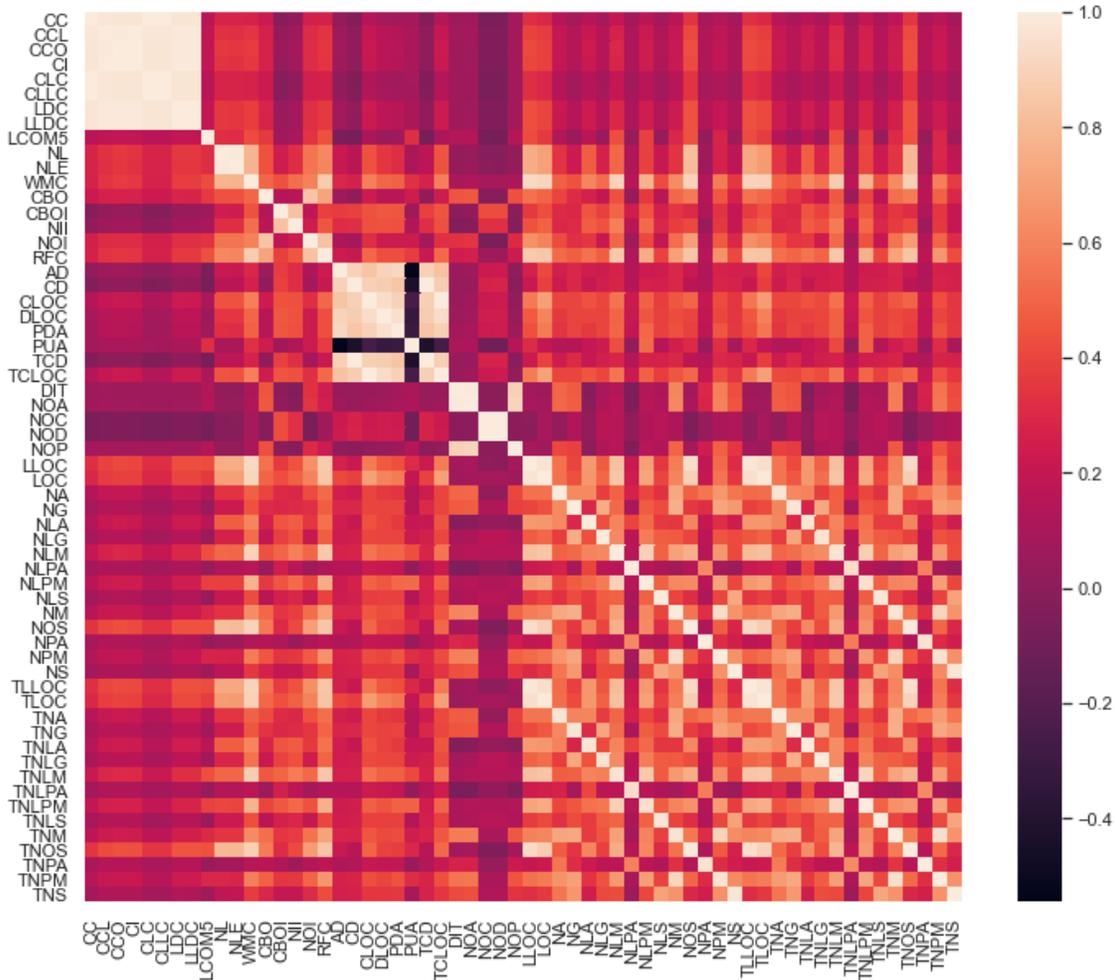


Fig. 4. A Heat Map Showing the Spearman Correlation Coefficients between the Software Metrics Existing in the used Dataset.

Support Vector Machine (SVM): SVM is a discriminative classifier algorithm that separates data samples into (generally) two different classes. The data samples are represented on 2-dimensional space and SVM tries to find an optimal hyperplane that divides the 2-dimensional space into two parts such that the data samples of each class reside on one part.

K-Nearest Neighbor (KNN): KNN is simple algorithm that solves the machine leaning classification problems by finding the k nearest neighbours (which has been previously classified) to the data point to be classified. Then KNN classifies the data point to the most frequent class in k nearest neighbours.

Naive Bayes (NB): NB a simple technique that is used to build a probabilistic classifier based on Bayes' theorem. NB classifiers assume that the input features are statistically independent. Thus, each feature contributes independently to the probability that an instance data belongs to a certain class regardless the correlation between the considered feature and other features.

Decision Tree (DT): DT is a learning technique that constructs a decision tree model for classification problems based on the given dataset. Each internal node in the constructed tree

symbolizes a test condition on a feature, each branch denotes an output of a test condition, and each leaf node denotes a label (or a decision).

Random Forest (RF): RF is an ensemble learning method that uses a set of unpruned decision trees for classification. The decision trees are constructed based samples of the dataset. During the classification of a data sample, the data sample is given to each decision tree and the class label of the instance is determined by taking the mode of the outputs of the decision trees.

Bagging: Bagging (aka Bootstrap aggregating) is an ensemble learning method that aims to improve the accuracy of machine learning algorithms most commonly decision trees. It reduces the variance of a model and helps to avoid the overfitting of data. It creates n subdatasets each of which contains a subset of features and data samples that are randomly selected with replacement from the original training dataset. The n subdatasets are used in parallel to construct n base (or weak) prediction models. The label class predicted by the majority of the base models is chosen to be the output of the bagging classifier.

AdaBoost: AdaBoost (stands for Adaptive Boosting) is an

ensemble learning method that is used to combine multiple weak classifiers into a single strong classifier for the purpose of improving the accuracy and performance of the weak classifiers. The weak classifiers in Adaboost are usually decision stumps which are decision tree with just one node (the root) and two leaves. The weak classifiers are trained sequentially. Samples misclassified by a weak classifier are given more weight in subsequent classifiers. Also each weak classifier is given a weight according to its accuracy. The final output of Adaboost classifier is the weighted sum of the outputs of the weak classifiers.

G. Performance Evaluation

Three performance metrics are used to evaluate the performance of the constructed models including accuracy (ACC), F-measure (F1), and Area under the ROC Curve (AUC). These metrics are widely used in the literature to evaluate SBP models.

Accuracy measures the fraction of the predictions that are classified correctly by a model. The value of accuracy ranges from 0 to 1 where a higher value means better accuracy. It is calculated for a binary classification according to the following equation:

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \quad (2)$$

where TP = True Positives, TN = True Negatives, FP = False Positives, and FN = False Negatives.

The F-measure is a commonly used performance metric (especially in case of the existence of an imbalanced classification problem) that considers both the precision and recall of a model. It is the harmonic mean of the precision and recall and its value ranges from 0 to 1 where a large value means a better performance. The following equation is used to compute the F-measure:

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall} \quad (3)$$

where the Precision and Recall are calculated as follows:

$$Precision = \frac{TP}{TP + FP} \quad (4)$$

$$Recall = \frac{TP}{TP + FN} \quad (5)$$

AUC is a performance metric for classification problems across all the various classification thresholds. The value of AUC lies between the closed interval [0, 1] where a larger value of AUC means better performance for a classifier. AUC is calculated based on the following equation:

$$AUC = \frac{\sum rank(All\ Positive\ Samples) - \frac{X(X+1)}{2}}{P * N} \quad (6)$$

Where P and N represent the number of positive and negative samples, respectively.

H. Tools

All the experiments in this study are implemented using the open source scikit-learn tools [26]. They are simple and efficient tools used to implement machine learning techniques in Python.

IV. RESULTS AND DISCUSSION

Following the research framework shown in Fig. 2, eight SBP models were constructed using the considered machine learning techniques and the used bug dataset. Out of the 60 features (Software metrics), 25 features were selected to construct the models after the application of the correlation-based filter-subset feature selection with best first search. The results of the considered performance metrics for each constructed model are given in Fig. 5.

The highest accuracy value is 0.82 which is achieved by LR, SVM, and RF classifiers. The accuracy values of NB, Adaboost, and KNN are 0.81, 0.8, and 0.79, respectively. The Bagging classifier has an accuracy of 0.76. DT has the lowest accuracy value which is 0.7.

LR, SVM, and RF classifiers attained the highest AUC value (0.77). NB achieved the second largest value of AUC (0.76). Adaboost has a value of 0.73 for AUC. The values of AUC for KNN, Bagging, and DT are 0.67, 0.62, and 0.53, respectively, which are relatively low compared to the values of AUC for other classifiers.

For the F1 values, LR and NB have the highest value (0.77). SVM, RF, and Adaboost attained the second largest value of F1 (0.76). KNN, Bagging, and DT have F1 values of 0.75, 0.73, and 0.7, respectively.

Answering RQ: From the results depicted in Fig. 5, it can be said that LR is the most effective machine learning technique (in terms of performance metrics) for bug prediction in classes based on the Unified Bug Dataset as it achieved the highest values of accuracy, AUC, and F1 measure. However, the performance of LR is not significantly better than the performance of the other considered techniques. In fact, some of the other classifiers have exactly the same performance of LR (for some of the used performance metrics) such as SVM and RF for the accuracy and AUC metrics and NB for the F1 metric.

V. THREATS TO VALIDITY

There are several issues that may impact the results of this study and limit their generality.

The quality of the bug dataset used to build the SBP models was not evaluated in this paper. The values of the software metrics and the bug information were used in all the experiments conducted in this study without verification or validation. However, the dataset was extensively evaluated and validated in [10] and it has been used in other studies in the literature (e.g. [11])

Software metrics included in the used dataset are not the only factors that can have impact on software bug proneness. Other factors such as the experience of software engineers involved in the development process of a software unit (e.g., class) and development environment can also make a software

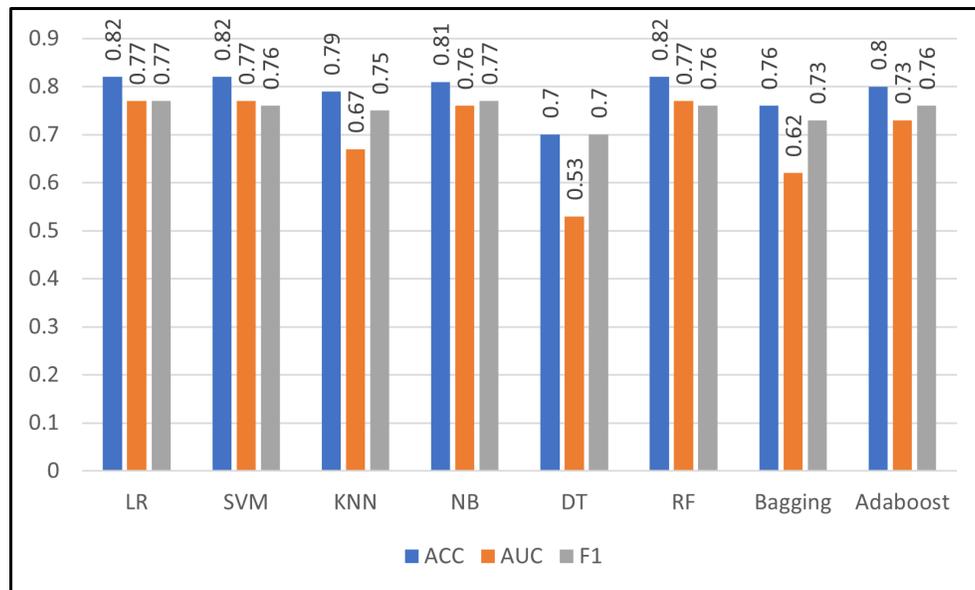


Fig. 5. The Average Results of the used Performance Metrics for the Constructed Models.

unit to be bug prone. However, these factors are out of the scope of this study and the aim of the study is to compare the performance of well-known machine learning techniques based on the used bug dataset which only includes software metrics and bug information.

The performance of the SBP models constructed in this study were evaluated based on only accuracy, F-1, and AUC. There are several other well-known performance metrics which were not used in this study such as Precision, Recall, Balance, and G-mean. However, there is no previous study on SBP that has used all the existing performance metrics to evaluate SBP models. In this study, the accuracy was used because it is one of the most commonly used metric to evaluate the performance of prediction models. In addition, the F-1 and AUC metrics were used in this study because the considered dataset is greatly imbalanced and these two metrics are good performance metrics for evaluating models constructed based on a biased dataset.

VI. CONCLUSION AND FUTURE WORK

In this study, the performance of 8 widely used machine learning techniques were investigated. A set of experiments were conducted using the Unified Bug Dataset which includes bug information for 47,618 classes. LR was found to be the most effective technique for predicting buggy classes. It attained 0.82, 0.77, and 0.77 for the accuracy, AUC, and F1, respectively.

The correlation-based filter-subset with best first search was the only feature selection technique applied in this study. There are many other feature selection and transformation techniques that have been introduced in the literature. A future work can extend this study by comparing the performance the considered 8 machine learning techniques when applying different feature selection and transformation techniques.

REFERENCES

- [1] A. O. Balogun, S. Basri, L. F. Capretz, S. Mahamad, A. A. Imam, M. A. Almomani, V. E. Adeyemo, A. K. Alazzawi, A. O. Bajeh, and G. Kumar, "Software defect prediction using wrapper feature selection based on dynamic re-reranking strategy," *Symmetry*, vol. 13, no. 11, p. 2166, 2021.
- [2] V. Nguyen, V. Pham, and V. Lam, "qestimation: a process for estimating size and effort of software testing," in *Proceedings of the 2013 International Conference on Software and System Process*, 2013, pp. 20–28.
- [3] H. Ohtera and S. Yamada, "Optimal allocation and control problems for software-testing resources," *IEEE Transactions on Reliability*, vol. 39, no. 2, pp. 171–176, 1990.
- [4] H. Tong, B. Liu, and S. Wang, "Software defect prediction using stacked denoising autoencoders and two-stage ensemble learning," *Information and Software Technology*, vol. 96, pp. 94–111, 2018.
- [5] V. Basili, L. Briand, and W. Melo, "A validation of object-oriented design metrics as quality indicators," *IEEE Transactions on Software Engineering*, vol. 22, no. 10, pp. 751–761, 1996.
- [6] X. Yang, H. Yu, G. Fan, and K. Yang, "A differential evolution-based approach for effort-aware just-in-time software defect prediction," in *Proceedings of the 1st ACM SIGSOFT International Workshop on Representation Learning for Software Engineering and Program Languages*, 2020, pp. 13–16.
- [7] D. Gray, D. Bowes, N. Davey, Y. Sun, and B. Christianson, "Using the support vector machine as a classification method for software defect prediction with static code metrics," in *International Conference on Engineering Applications of Neural Networks*. Springer, 2009, pp. 223–234.
- [8] H. Osman, M. Ghafari, and O. Nierstrasz, "Hyperparameter optimization to improve bug prediction accuracy," in *2017 IEEE Workshop on Machine Learning Techniques for Software Quality Evaluation (MaLTesQuE)*. IEEE, 2017, pp. 33–38.
- [9] [Online]. Available: <http://promise.site.uottawa.ca/SERespository/datasets-page.html>
- [10] R. Ferenc, Z. Tóth, G. Ladányi, I. Siket, and T. Gyimóthy, "A public unified bug dataset for java and its assessment regarding metrics and bug prediction," *Software Quality Journal*, vol. 28, no. 4, pp. 1447–1506, 2020.
- [11] Z. J. Szamosvölgyi, E. T. Váradi, Z. Tóth, J. Jász, and R. Ferenc, "Assessing ensemble learning techniques in bug prediction," in *International Conference on Computational Science and Its Applications*. Springer, 2021, pp. 368–381.

- [12] T. McCabe, "A complexity measure," *IEEE Transactions on Software Engineering*, vol. SE-2, no. 4, pp. 308–320, 1976.
- [13] M. H. Halstead, *Elements of Software Science (Operating and programming systems series)*. Elsevier Science Inc., 1977.
- [14] S. Chidamber and C. Kemerer, "A metrics suite for object oriented design," *IEEE Transactions on Software Engineering*, vol. 20, no. 6, pp. 476–493, 1994.
- [15] T. Wang and W.-h. Li, "Naive bayes software defect prediction model," in *2010 International Conference on Computational Intelligence and Software Engineering*. Ieee, 2010, pp. 1–4.
- [16] P. D. Singh and A. Chug, "Software defect prediction analysis using machine learning algorithms," in *2017 7th International Conference on Cloud Computing, Data Science & Engineering-Confluence*. IEEE, 2017, pp. 775–781.
- [17] F. Matloob, T. M. Ghazal, N. Taleb, S. Aftab, M. Ahmad, M. A. Khan, S. Abbas, and T. R. Soomro, "Software defect prediction using ensemble learning: A systematic literature review," *IEEE Access*, 2021.
- [18] J. Petrić, D. Bowes, T. Hall, B. Christianson, and N. Baddoo, "The jinx on the nasa software defect data sets," in *Proceedings of the 20th International Conference on Evaluation and Assessment in Software Engineering*, 2016, pp. 1–5.
- [19] Z. Xu, J. Liu, Z. Yang, G. An, and X. Jia, "The impact of feature selection on defect prediction performance: An empirical comparison," in *2016 IEEE 27th International Symposium on Software Reliability Engineering (ISSRE)*. IEEE, 2016, pp. 309–320.
- [20] Sed-Inf-U-Szeged, "Sed-inf-u-szeged/openstaticanalyzer: Openstaticanalyzer is a source code analyzer tool, which can perform deep static analysis of the source code of complex systems." [Online]. Available: <https://github.com/sed-inf-u-szeged/OpenStaticAnalyzer>
- [21] M. Jureczko and L. Madeyski, "Towards identifying software project clusters with regard to defect prediction," in *Proceedings of the 6th international conference on predictive models in software engineering*, 2010, pp. 1–10.
- [22] T. Zimmermann, R. Premraj, and A. Zeller, "Predicting defects for eclipse," in *Third International Workshop on Predictor Models in Software Engineering (PROMISE'07: ICSE Workshops 2007)*. IEEE, 2007, pp. 9–9.
- [23] M. D'Ambros, M. Lanza, and R. Robbes, "An extensive comparison of bug prediction approaches," in *2010 7th IEEE working conference on mining software repositories (MSR 2010)*. IEEE, 2010, pp. 31–41.
- [24] T. Hall, M. Zhang, D. Bowes, and Y. Sun, "Some code smells have a significant but small effect on faults," *ACM Transactions on Software Engineering and Methodology (TOSEM)*, vol. 23, no. 4, pp. 1–39, 2014.
- [25] Z. Tóth, P. Gyimesi, and R. Ferenc, "A public bug database of github projects and its application in bug prediction," in *International Conference on Computational Science and Its Applications*. Springer, 2016, pp. 625–638.
- [26] "Sklearn.preprocessing.minmaxscaler." [Online]. Available: <https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.MinMaxScaler.html>
- [27] B. Ghotra, S. McIntosh, and A. E. Hassan, "A large-scale study of the impact of feature selection techniques on defect classification models," in *2017 IEEE/ACM 14th International Conference on Mining Software Repositories (MSR)*. IEEE, 2017, pp. 146–157.
- [28] Z. Mahmood, D. Bowes, P. C. Lane, and T. Hall, "What is the impact of imbalance on software defect prediction performance?" in *Proceedings of the 11th International Conference on Predictive Models and Data Analytics in Software Engineering*, 2015, pp. 1–4.
- [29] J. Chen, S. Liu, W. Liu, X. Chen, Q. Gu, and D. Chen, "A two-stage data preprocessing approach for software fault prediction," in *2014 Eighth International Conference on Software Security and Reliability (SERE)*. IEEE, 2014, pp. 20–29.
- [30] M. Liu, L. Miao, and D. Zhang, "Two-stage cost-sensitive learning for software defect prediction," *IEEE Transactions on Reliability*, vol. 63, no. 2, pp. 676–686, 2014.
- [31] P. He, B. Li, X. Liu, J. Chen, and Y. Ma, "An empirical study on software defect prediction with a simplified metric set," *Information and Software Technology*, vol. 59, pp. 170–190, 2015.

Transformer-based Models for Arabic Online Handwriting Recognition

Fakhraddin Alwajih^{1,2} *, Eman Badr^{1,3}, and Sherif Abdou¹

Department of Information Technology, Cairo University, Giza, Egypt¹

Department of Computer Science and Information Technology, Ibb University, Ibb, Yemen²

University of Science and Technology, Zewail City of Science, Technology and Innovation, Giza, Egypt³

Abstract—Transformer neural networks have increasingly become the neural network design of choice, having recently been shown to outperform state-of-the-art end-to-end (E2E) recurrent neural networks (RNNs). Transformers utilize a self-attention mechanism to relate input frames and extract more expressive sequence representations. Transformers also provide parallelism computation and the ability to capture long dependencies in contexts over RNNs. This work introduces a transformer-based model for the online handwriting recognition (OnHWR) task. As the transformer follows encoder-decoder architecture, we investigated the self-attention encoder (SAE) with two different decoders: a self-attention decoder (SAD) and a connectionist temporal classification (CTC) decoder. The proposed models can recognize complete sentences without the need to integrate with external language modules. We tested our proposed models against two Arabic online handwriting datasets: Online-KHATT and CHAW. On evaluation, SAE-SAD architecture performed better than SAE-CTC architecture. The SAE-SAD model achieved a 5% character error rate (CER) and an 18% word error rate (WER) against the CHAW dataset, and a 22% CER and a 56% WER against the Online-KHATT dataset. The SAE-SAD model showed significant improvements over existing models of the Arabic OnHWR.

Keywords—Self attention; Transformer; deep Learning; connectionist temporal classification; convolutional neural networks; Arabic online handwriting recognition

I. INTRODUCTION

OnHWR is essentially a task of converting digital input handwriting into digital text. Handwriting recognition can be classified into two main categories based upon input data: online and offline handwriting recognition. In online handwriting, data is represented as a series of points with the precision of other information, such as timestamps, dependent upon the capabilities of the input device. In offline handwriting recognition, data is represented as images scanned from documents.

In recent years, OnHWR has attained increased importance concomitant with rapid developments in related hardware and software. Most current communication software supports notetaking and writing on boards using online handwriting as both a communication media and a vehicle of computer-aided education. In the rising markets, greater access to computing devices has allowed ever-increasing populations to connect across the internet, with many depending solely on mobile devices with touchscreens. Handheld devices with styluses are becoming more widely available and used in many domains.

Concomitantly, there have been tremendous advances in prime technologies of deep learning and natural language processing (NLP) algorithms. Such advances have led, in turn, to considerable progress in the field of OnHWR. The Arabic language is spoken by around half a billion people around the world. A number of other languages, including, Urdu, Persian, Kurdish, and Pashto adopted and use Arabic script. Arabic is a 'right to left' language in its written form. It consists of 28 letters, 10 digits as well as a number of punctuation marks. Each Arabic letter has four contextual forms, depending upon its position in a word: isolated, beginning, middle, and end position forms, as shown in Fig. 1. Arabic OnHWR is a challenging problem for multiple reasons. One reason is the existence of a wide range of variations in handwriting styles, in part due to the existence of multiple calligraphies in Arabic. There are eight basic calligraphies in Arabic script [1]. The tendency is to use a combination of these calligraphies when writing in Arabic. This further compounds the variations in styles of writing, thus adding to the challenges that would face the developer of an Arabic script recognition system. Compared to Latin and Chinese and other scripts, published work in the Arabic OnHWR field has to date been fairly limited.

OnHWR is a sequence-to-sequence (S2S) classification task. Input frames are ingested into the S2S model which in turn generates text. Recent advances in S2S models have shown their reliability solve complex NLP tasks such as translation [2] and automatic speech recognition (ASR) [3]. Additionally, the performance of OnHWR systems has improved with the advent of deep learning models including convolutional neural network (CNN) [4] and long short-term memory (LSTM) [5], [6].

Recently, E2E OnHWR systems have achieved remarkable performance, with input handwriting features being mapped directly to an output sequence of letters or tokens. In E2E systems, all components are trained and optimized jointly, thus reducing the complexity of the system and minimizing error propagation between components compared to conventional hybrid systems. Using CTC, E2E modeling has been utilized for handwriting recognition tasks as well as attention-based encoder-decoder systems designed for mathematical expression recognition tasks [7], [8]. Moreover, E2E has been incorporated with external language models (LM), effectively boosting performance [5]. In general, the competitive performance obtained by E2E models and their simplicity facilitate the building of state-of-the-art OnHWR systems. In this work, we explore building an E2E OnHWR system based on self-

*Corresponding authors.

Contextual forms				Name
Isolated	End	Middle	Beginning	
ا	ا	-	-	Alif
ب	ب	ب	ب	Beh
ت	ت	ت	ت	Teh
ث	ث	ث	ث	Theh
ج	ج	ج	ج	Jeem
ح	ح	ح	ح	Hah
خ	خ	خ	خ	Khah
د	د	-	-	Dal
ذ	ذ	-	-	Thal
ر	ر	-	-	Reh
ز	ز	-	-	Zain
س	س	س	س	Seen
ش	ش	ش	ش	Sheen
ص	ص	ص	ص	Sad
ض	ض	ض	ض	Dad
ط	ط	ط	ط	Tah
ظ	ظ	ظ	ظ	Zah
ع	ع	ع	ع	Ain
غ	غ	غ	غ	Ghain
ف	ف	ف	ف	Feh
ق	ق	ق	ق	Qaf
ك	ك	ك	ك	Kaf
ل	ل	ل	ل	Lam
م	م	م	م	Meem
ن	ن	ن	ن	Noon
ه	ه	ه	ه	Heh
و	و	-	-	Waw
ي	ي	ي	ي	Yeh

Fig. 1. Arabic Letters and their Contextual Forms.

attention models.

RNNs have been adopted for sequence modeling and have provided remarkable accuracy in multiple NLP tasks [9], [2], [10]. RNNs have been extensively utilized in OnHWR, including in the build-up of LSTM and gated recurrent units (GRU). In RNN, each hidden state depends on the previous one which makes parallelizing computations of RNNs difficult. Additionally, the hidden states are condensed into a fixed-length vector which introduces a 'bottleneck' making capturing long dependencies difficult as well [10].

As alternatives to RNNs, transformer-based models [11] have recently yielded outstanding results, achieving state-of-the-art performance in a variety of NLP tasks, including text and image-related tasks, and ASR [12], [11], [13]. Transformers rely on a self-attention mechanism, which extracts a more representative sequence by relating all position pairs of an input sequence. The self-attention mechanism offers two

attractive features compared with RNNs: (1) computations can be parallelized and carried out efficiently through batched tensor operations, and (2) self-attention allows direct connection for long-range and short-range dependencies without propagating contextual information between intermediate hidden states (as in case of RNNs) [11]. In addition to self-attention, the transformer model utilizes multi-head attention (MHA) in order to learn different representations in one instant. As with RNNs attention-based models, transformers are architecturally designed as encoder-decoder models, with both the encoder and decoder containing stacked self-attention networks (SANs) on top of each other. The cross-attention mechanism is used to bridge between the encoder and the decoder. The successes of transformer models have inspired this work in which self-attention was applied to an OnHWR task.

In this paper, we introduce transformer-based models OnHWR for Arabic script. The proposed models can transform a full-sentence handwriting input sequence into the corresponding letter sequence. Basically, we applied CNN layers to subsample input sequence features (via convolution strides) and process local relationships between handwriting frames of the input sequence. The output is added to positional embedding output to maintain input orders and then fed into the self-attention encoder (SAE). For the decoder, we employed two decoders: the self-attention decoder (SAD) and the CTC decoder. The proposed models were trained and evaluated against two datasets: Online-KHATT dataset [14] and CHWA dataset [15]. To the best of our knowledge, there has been no prior work on OnHWR that has proposed or applied self-attention models. As far as we are aware, this is the first attempt to apply the transformer model to an OnHWR task. Our results show that our proposed SAE-SED model can outperform existing RNNs models.

The main contributions can be summarized as follow:

- We introduce a new self-attention-based non-recurrent neural network models for OnHWR task.
- Two architectures have been developed in the decoding stage for the transformer: a SAD decoder and a CTC decoder.
- The proposed models have been evaluated against a full sentence (OnKHATT) [14] and a word-based (CHAW) Arabic dataset [15]. Results were compared with existing models, with our model evidently outperforming these models.

The rest of this paper is structured as follows: Section II details related work previously conducted on OnHWR. In Section III we layout the architecture of the transformer we designed for the OnHWR task. Then, experimental results are presented in Section IV. Lastly, Section V details our conclusions and recommendations for possible future work.

II. RELATED WORK

OnHWR data have a temporal structure and can be represented as a sequence of geometrical features vectors over time. OnHWR relies on sequence modeling, including statistical modeling. Previously, hidden Markov models (HMMs) have been reportedly utilized to model online handwriting in multiple published works. In [16], the HMM was designed

to model stroke segments as handwriting model units. In their model, letter models are subsequently formed by concatenating model units as defined in a pronunciation dictionary. Letter models are integrated as word sequence probabilities to form a stochastic language model. In [17], The researchers integrated Gaussian mixture models (GMMs) with the HMMs as continuous HMMs, using the GMMs to estimate observation probability distributions emitted by HMM states. Hybrid HMMs with feed-forward neural networks (NNs) were also part of their design [18]. Authors in [19] integrated a time-delay neural network (TDNN) with an HMMs into a single architecture, combining the recognition and segmentation phases into a single hybrid architecture. In their model, this hybrid architecture was intended to utilize the power of TDNN in the recognition and power of HMMs in the segmentation.

Traditional approaches involve multiple components that are separately trained and optimized, introducing suboptimality. On the other hand, deep learning models work on feature representation by learning discriminative representation from the raw data, thus providing an E2E solution for concomitant training of OnHWR system components jointly. One of the first works used implicit segmentation that was to be trained jointly with the recognition phase in [20]. In [20], the connectionist temporal classification (CTC) loss was introduced as an objective to map input frames into letters and optimize recognition jointly with LSTM.

Deep CNN was also utilized by [21]. In this study, the authors integrated CNN with domain-specific technologies to form an integrated network to improve performance. The efficacy of a combination of a CNN, RNN, and CTC was also investigated by [22]. They placed CNN layers at the front in order to support features representation. Next, they added LSTM to model the sequence of OnHWR along with a CTC to optimize the integrated network in an E2E manner. Authors investigated handcraft features and raw data ingested to CNN and reported that the proposed model had performed better with handcraft features. Furthermore, in [1], CNN-BiLSTM-CTC architecture was used to design an Arabic OnHWR model.

Recent work by Google investigated a model consisting of bidirectional LSTM (BiLSTM) with CTC [5]. In this work, the authors utilized a BiLSTM encoder trained using CTC loss. In decoding, they used different scoring LMs to incorporate prior knowledge about the underlying language and decode the output of the RNN encoder. GRU was employed in S2S with attention architecture and used in recognition tasks of online handwriting data of mathematical expression [8], [23], which was originally used in neural machine translation (NMT) by [24]. In [25], authors utilized attention encoder-decoder to recognize unconstrained Vietnamese Handwriting. The encoder was fronted with a CNN to extract invariant features and a BiLSTM to encode the output of CNN. The decoder was composed of BiLSTM layers with attention incorporated with encoders in order to generate text output. In [26], an edge graph attention network (EGAT) was proposed as a model that would perform stroke recognition. Stroke classification was formulated as node classification in a graph neural network (GNN).

In Arabic OnHWR, the line of work simply flows the Latin OnHWR workflow [27], [28]. As with traditional OnHWR sys-

tems, HMMs was utilized in many works for Arabic OnHWR [29], [30], [31], [32], [33]. Hybrid NN/HMMs were investigated in [34] and a DNN/HMMs model was tested by [15]. CTC based models were employed in several works for Arabic OnHWR [35], [36], [37], [1]. Most of the aforementioned studies targeting Arabic OnHWR tested their models against word-based datasets, with the exception of our previous work [35], [1] in which we tested our models against both sentence-based and word-based datasets. In our previous work [35], we proposed an E2E BiLSTM-CTC model and incorporated LM with outputs of RNNs to boost the performance of the system. More recently, we developed a writer adaptation method that utilized an E2E CNN-BiLSTM-CTC model [1]. In the current work, we did not integrate with any external module, and we evaluated our work against the CHWA and Online-KHATT datasets.

Variations can be reduced by normalization preprocessing steps. Normalization acts by reducing geometric variants in order to facilitate extracting features that are relevant to recognition. In the literature on OnHWR, multiple normalization methods, including slant correction, smoothing using a Gaussian filter, and resampling, have been proposed and tested against OnHWR data. The most comprehensive preprocessing steps were detailed by [38].

Features extraction refers to the process of extracting a meaningful set of features from raw data to be ingested and eased in the recognition phase. In OnHWR, traditional features can be classified into local features per point and global features per stroke or character [38]. Recently, as deep learning helped perfect features representation, the need for handcraft features with learnable features representation was eliminated in such areas as NLP [39], ASR [3], and computer vision [40]. Two recent works in which the authors used deep learning for features representation are [5], [21]. Despite its advantages, deep learning needs large-scale datasets to learn features representation and OnHWR datasets are rare and limited in size.

To summarize, state-of-the-art OnHWR models based on deep recurrent networks have begun to achieve remarkable recognition results, although training is computationally expensive and takes a long time to converge. Furthermore, the problem with pure RNN methods is that information may be forgotten during the encoding process, thus degrading overall model performance. In this work, we propose the usage of transformer-based models for the OnHWR task for the first time with no-recurrent design. A single, unified E2E architecture capable of recognizing full sentences from input online handwriting without the need for predetermined lexicons or language models.

III. TRANSFORMER FOR ONHWR

Typically, the OnHWR is an S2S task in which the lengths of input and output can differ. In our framework, the architecture of the transformer is based on an encoder-decoder structure. Given the handwriting input sequence $X = (x_1, x_2, \dots, x_{T^{in}})$, $x_i \in R^{d_{in}}$ where T^{in} is the length of input sequence and d_{in} is the number of features. Before feeding the input into the encoder, we prepended the encoder with CNN layers to extract better representative handwriting features and

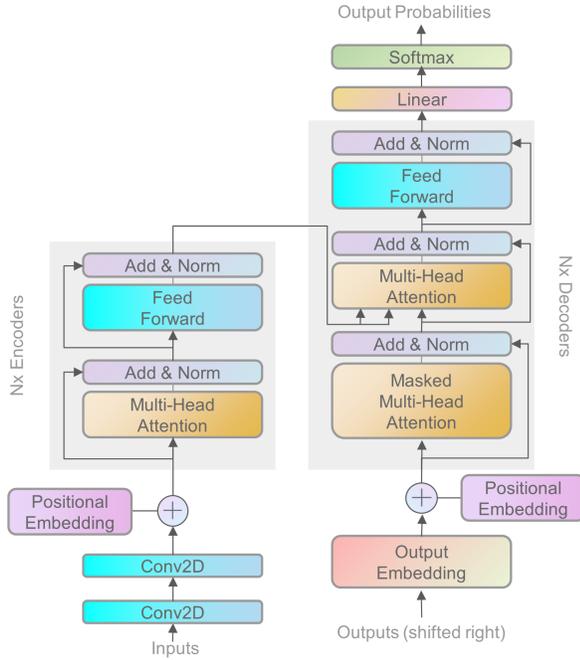


Fig. 2. The Architecture of OnHWR Transformer.

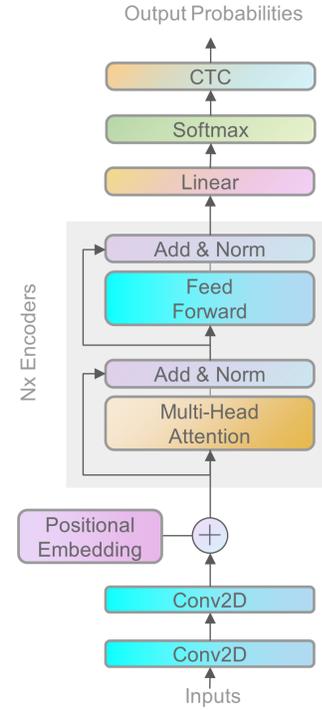


Fig. 3. Self Attention Encoder (SAE) with CTC.

perform subsampling on the input handwriting frames. Then, we applied positional embedding to the output of CNN layers $X = (x_1, x_2, \dots, x_T), x_i \in R^{d_{model}}$, where $T = |X|$ after subsampling. Positional embedding affords the input sequence a perception of order. The encoder was designed to encode the input sequence $X = (x_1, \dots, x_T), x_i \in R^{d_{model}}$ and generate intermediate updated representation using a self-attention mechanism $h = (h_1, h_2, \dots, h_T), h_i \in R^{d_{model}}$. The decoder transduces the output sequence using autoregressive approach. Given h representation and previously emitted characters of the decoder outputs to that point $y_{i-1} = (y_1, \dots, y_{i-1})$, the decoder computes the next character y_i . This procedure is repeated until the end of the sentence token is emitted as shown in Figure 2.

We also examined using a CTC decoder instead of the transformer decoder in which h representation would be ingested directly into the linear output layer. The output $y = (y_1, \dots, y_L)$ with length L is emitted by CTC decoder at once as shown in Fig. 3.

A. Self-Attention

The transformer-based models are built on a new concept of self-attention as an extension of attention introduced in S2S [24], [2]. Self-attention is a mechanism to compute updated representations for each sequence element in parallel. The attention mechanism would allow each representation to differentially consider the representations in every other position in a sense, and the communication paths would have the same length for all pairs of elements. The attention mechanism consists of a query matrix Q , a key matrix K , and a value V matrix. The basic idea is that a query vector would be compared to a set of key vectors to determine their

rapport. Each key vector comes paired with a value vector. The greater the rapport of a given key with the query the greater influence the corresponding value would have on the output of the attention mechanism. Transformer employs **scaled dot product attention** to map a query with a series of key-value pairs to output using the following equation:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (1)$$

where $Q \in R^{M \times d_k}$ and $K, V \in R^{N \times d_k}$ denote queries, keys and values in the matrix form, M and N are the number of queries and key-value pairs, and d_k is the dimension of representation. Scaling by factor $\sqrt{d_k}$ is done to prevent extremely small gradients.

B. Multi Head Attention (MHA)

Using a single attention head, the linear combination of value vectors leads to an averaging outcome that restricts the resolution of the learned representations. Therefore, the authors propose using multiple attention heads that can simultaneously learn different representations to alleviate this. MHA is computed as follows:

$$\begin{aligned} \text{MultiHead}(Q, K, V) &= \text{Concat}(\text{head}_1, \dots, \text{head}_h)W^O \\ \text{where } \text{head}_i &= \text{Attention}(QW_i^Q, KW_i^K, VW_i^V) \\ W_i^Q &\in R^{d_{model} \times d_k}, \\ W_i^K &\in R^{d_{model} \times d_k}, \\ W_i^V &\in R^{d_{model} \times d_v}, \\ W_i^O &\in R^{hd_v \times d_{model}} \end{aligned} \quad (2)$$

First, the inputs: query matrix Q , key matrix K and value matrix V are linearly projected using W^Q , W^K and W^V . The projected query QW^Q , key KW^K and value VW^V are split into h heads. Scaled dot-product attention is computed for each head i . The independent attention head computed are then concatenated and linearly projected using W^O .

C. Self-Attention Encoder (SAE)

Instead of positional encoding proposed in the original paper, we adopted learnable positional embedding [41]. The positional embedding has the same dimensionality as the input embedding, and we summed them together before feeding them to the encoder. MHA is the first of two sub-layers of an encoder layer. After each sublayer, both residual connection and layer normalization were applied. The residual connection adds a copy of the input to the output, which means the input representations before an MHA block are added to the output representations. Then layer normalization takes the input vectors and essentially normalizes each one individually to have zero mean and variance. This is done to assure training stability. The second sub-layer is a **position-wise feed-forward network**, composed of a simple network of two fully connected layers with value activation between them to each input representation as follows:

$$FFN(x) = \max(0, xW_1 + b_1)W_2 + b_2 \quad (3)$$

After the second sub-layer, we again applied a residual connection and layer normalization, thus completing one encoder layer. The aforementioned layers can then be stacked up N times to form the full encoder.

D. Self Attention Decoder (SAD)

The design of the self-attention decoder mimicked that of the aforementioned encoder architecture, except that it is composed of two MHA layers. The first MHA layer applies attention to outputs generated by the decoder up to a point. The first layer is masked to avoid attending to future positions, while the second MHA layer applies attention to the encoder outputs.

E. The CTC Decoder

CTC objective loss was described by [7], [20]. CTC directly estimates prediction labels in E2E models without the need for explicit segmentation or alignment between input frames and output labels. As with the RNN encoder [35], the encoder (SAE) outputs sequences with the same length of input sequence frame length of the input. CTC manages this condition by introducing an additional blank label b symbol to the target labels and allowing repetition of labels or by adding banks across frames to match the lengths of input frames. Given input handwriting frames $\mathbf{x} = (x_1, \dots, x_T)$, where $T = |\mathbf{x}|$ and $x_t \in R^{\text{d}_{\text{model}}}$ and output sequence labels $\mathbf{y} = (y_1, \dots, y_L)$, where $L = |\mathbf{y}|$ and $y_l \in Z$ and Z denote the (finite) label alphabet, the encoder (SAE) generates posteriors $P(\mathbf{y}|\mathbf{x})$ as follows:

$$P(\mathbf{y}|\mathbf{x}) = \sum_{\hat{\mathbf{y}} \in H_{CTC}(\mathbf{x}, \mathbf{y})} \prod_{t=1}^T P(\hat{y}_t | x_1, \dots, x_T)$$

where $\hat{\mathbf{y}} = (\hat{y}_1, \dots, \hat{y}_T) \in H_{CTC}(\mathbf{x}, \mathbf{y}) \subset \{Z \cup b\}^T$ harmonizes to any possible paths under the condition that $\hat{\mathbf{y}}$ yields \mathbf{y} after dropping the blank symbols b and repeated successive symbols of $\hat{\mathbf{y}}$. The CTC loss assumes that each label in the output sequence is conditionally independent given the input handwriting sequence.

IV. EXPERIMENTS AND RESULTS

A. Datasets

We tested our models against two open vocabulary datasets, the Online-KHATT and CHAW. The Online-KHATT dataset is an open vocabulary dataset collected by KFUPM [14]. It is comprised of 10,040 sentences of Arabic text written by 623 writers using Windows and Android run devices. Writers that contributed to the Online-KHATT dataset represent different ages, education levels, nationalities, genders, and handedness. This dataset consists of natural and unrestricted handwriting styles. The Online Arabic handwriting Cairo University dataset (CHAW) [15] is a word-level collection of Arabic writing. CHAW was collected using Android Samsung tablets. It consists of 18k of distinct words within a total of 192k samples. These samples are split into a training set, consisting of 17k unique words within a total of 180k samples, and a testing set, consisting of 500 unique words within 12k samples. A total of 1250 writers had contributed to this dataset. These writers are of varied ages, genders, and handedness.

B. Model Description

For input features, all preprocessing steps and features described in [17], [20] are used except delayed strokes representation features. The input sequence consists of a vector of 20 features per point. We normalized the input samples using z-score normalization before samples are fed into the models. For output, we adopted 160-character units, including 28 Arabic characters and their variations at different positions within a word, numbers, blank, punctuations, a start of sentence label (SOS) and end of sentence label (EOS).

We placed 2 CNN layers for the purposes of handwriting feature embedding. To stabilize training, we applied batch normalization (BN) [10] after each CNN layer, followed by ReLU activation. In our models, CNN layers perform subsampling by time reduction of the input frame handwriting sequence and retaining more representative features.

For the SAE-SAD model shown in Fig. 2, we stacked 6 SAE encoders and only one SAD decoder layer. In addition, we used $h_{\text{heads}} = 4$ for MHA. A total of 256 units comprised the feed-forward sub-layers. For the SAE-CTC model shown in Fig. 3, we mimicked the structure of the SAE-SAD model, replacing the SAD layer with a CTC as described in Section III-E.

In the training phase, we used an Adam optimizer with a learning rate scheduling [11]. In cross-entropy loss, which is used to optimize the SAE-SAD model, we applied label smoothing with a plenty factor of 0.1 [42]. SAE-CTC model optimized, the entire model using CTC loss. To avoid overfitting during training, dropout, at a rate of 0.3, is used [43]. In the end, we averaged the parameters of models of the last five epochs [44].

TABLE I. THE CERs COMPARISONS OF DIFFERENT HYPERPARAMETERS COMBINATIONS FOR SAE-SAD MODEL.

# of encoders	# of decoder	h_{heads}	d_{ff}	CER [%]
4	4	2	128	34.77 %
4	4	4	256	28.64 %
6	4	4	256	23.16 %
6	2	4	256	21.35 %
6	1	4	256	20.03 %
6	1	4	512	25.88 %
8	1	4	256	23.47 %

TABLE II. COMPARING OUR MODELS TO OTHER HYBRID AND E2E SYSTEMS REPORTING ON ONLINE-KHATT AND CHAW.

Models	CHAW dataset		Online-KHATT dataset	
	CER [%]	WER [%]	CER [%]	WER [%]
DNN/HMMs [15]	-	25%	-	-
BiLSTM-CTC-LM [35]	4.08%	14.65%	12.24%	28.35%
CNN-BiLSTM-CTC [1]	9.43%	34.48%	18.49%	59.94%
SAE-CTC	10.83%	40.68%	23.89%	78.17%
SAE-SAD	5.70%	18.45%	22.88%	56.48%

C. Results

We used the standard matrices word error rate (WER) and character error rate (CER) to evaluate our experiment results. WER is calculated by summing up insertions, substitutions, and deletions present in recognized words divided by the length of words in the target sentence. CER is calculated in a similar fashion, this time focusing on characters instead of words.

To select the best hyperparameters for our proposed mod-

els, we ran multiple experiments of different hyperparameter combinations, varying the number of blocks in encoders, feed-forward units in the sub-layers of each block, number of heads h_{heads} for the encoder and number of blocks of the decoder in SAE-SAD. For the subsampling CNN module, we followed the architecture and hyperparameters in [1]. Table I shows different configurations we had tried for SAE and SAD with the CER on the validation set.

We trained the SAE-SAD model for 228 epochs and SAE-CTC model for 60 epochs. Training stopped when models started overfitting. We then selected the best model with the lowest CER on the validation set. Fig. 4, shows a comparison of validation loss and training loss for both SAE-SAD and SAE-CTC models, respectively. We also compared CER and WER on the validation dataset for both the SAE-SAD and SAE-CTC models. We trained all models using a GeForce GTX 1080 Ti, and we conducted all experiments using a Tensorflow [45]. At this scale, 228 epochs of SAE-SAD model run over 112 hours, whilst SAD-CTC model took over 30 hours. In Fig. 4, we see that the SAE-CTC model converges faster than the SAE-SAD model. However, the SAE-SAE model took more epochs to converge, and its WER was superior to that of the SAE-CTC model. We also found that CER was closer to WER in the case of the SAE-SAD model than in the case of the SAE-CTC model, indicating the SAE-SAE model to be capable of capturing words more accurately at a higher rate than the SAE-CTC model.

The online-KHATT dataset is challenging and contains sentence-based samples and a subset of segmented characters. All previous works, [46], [47], [48], [35], [1] conducted their experiments against the character set in Online- KHATT except

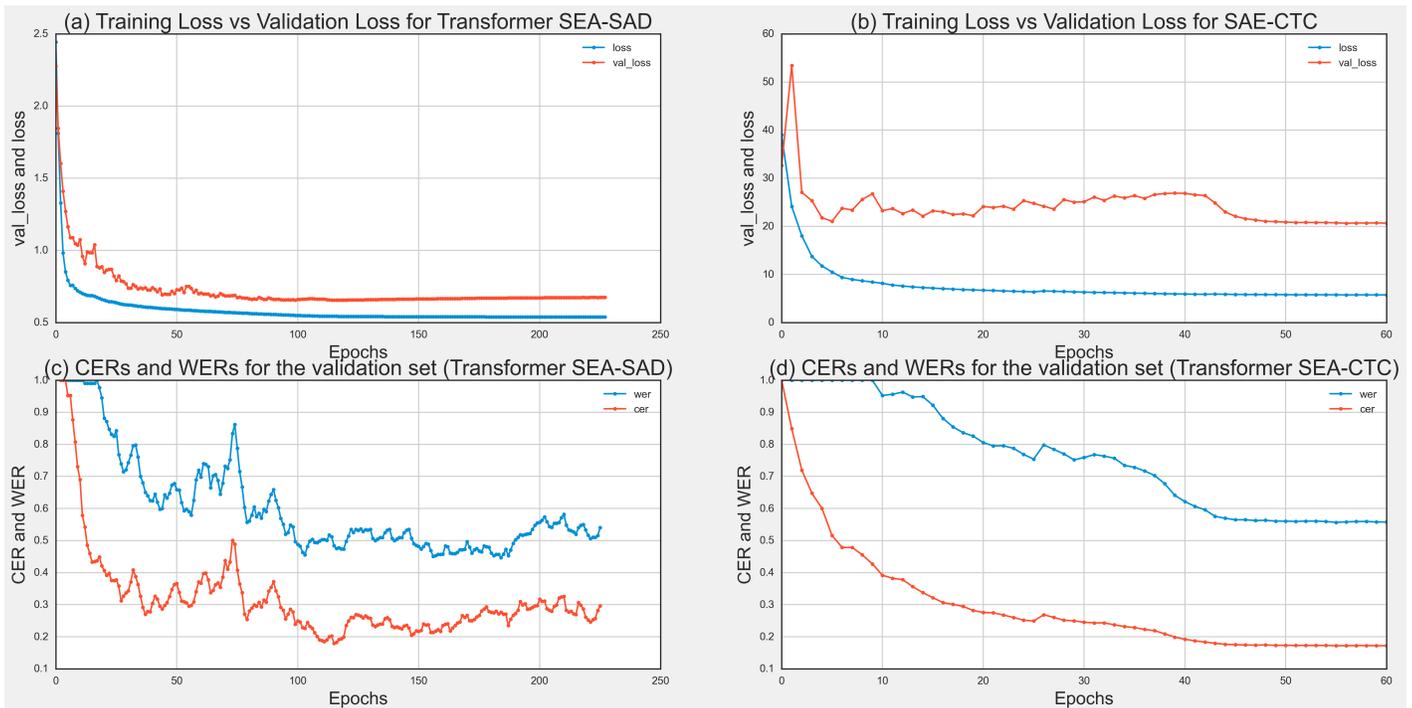


Fig. 4. (a) Training and Validation Losses for (a) SEA-SAD Model (b) SEA-CTC Model and CERs and WERs on Validation Dataset (c) SEA-SAD Model (d) SEA-CTC Model

[1], [35]. In our work, we compared our proposed models to existing systems that had tested their models against the full sentence-based set in the Online-KHATT dataset.

Table II shows the evaluation results on Online-KHATT and CHAW datasets. For the hybrid DNN/HMM-based approach in [15], the authors evaluated their work on the CHAW dataset, which is a word-based dataset. Furthermore, they integrated a dictionary with the model output to improve the results of the proposed approach. For the E2E system in our previous work [35], we incorporated n-gram LM to boost the result of the proposed approach, and we evaluated the proposed method on both Online-KHATT and CHAW datasets. Naturally, LMs boost the result of DNN models, and in this work, we have not incorporated any external LM or dictionary. Thus, this approach is not comparable to this work. The bottom row in Table II compares our previous CNN-BiLSTM-CTC [1] model with the proposed models since it does not integrate any external module. We compared our results with [1] results of the writer independent model as this result was achieved on the whole test dataset. The results show that SAE-SAD model outperforms our prior CNN-BiLSTM-CTC model [1]. Also, SAE-SAD outperforms SAE-CTC models. In addition, our proposed SAE-SAD performs better than the hybrid DNN/HMM model [15].

D. Discussion

Deep learning models learn to model discriminative features representation. As shown in Table I, deeper encoders perform better as we increase encoder layers. This is because each layer learns at a different level of abstraction for a given set of features. Multiple encoder layers are capable of generalization because each layer learns a different intermediate representation of raw data which helps at the classification level.

E2E CTC-based models are typically trained jointly using the loss CTC function. However, CTC-based models assume that relationships among produced labels from the CTC-based model are conditionally independent. Thus, such models cannot implicitly learn the LM from the training data. On the other hand, transformer-based models with a SAD decoder generate with each time step a label that is conditionally dependent on the previously generated ones. Consequently, they are capable of capturing the LM directly from training data. This would explain why the SAE-SAD model outperformed the SAE-CTC model, as shown in Table II. Also, we believe that SAE-SAD models could outperform traditional models that are integrated with external LMs in the presence of sufficient data. However, one advantage of the SAE-CTC model compared with the SAE-SAD model is its ability to generate the output labels in parallel at inference time.

CNN networks are widely used in transformer-based ASR models for down-sampling as well as providing positional encoding [13]. However, in the case of our OnHWR models, CNNs did not provide sufficient order information to the models other than that contributed through subsampling. Thus, we utilized positional embedding to add order sense to CNN outputs before feeding them into the encoder. The inability of CNN to provide sufficient order information may be due to the nature of handwriting data which contains delayed strokes, and

the limited nature of Arabic handwriting datasets. We found that adding learnable positional embedding made the model converge faster.

V. CONCLUSION

In this work, we have introduced self-attention based Arabic OnHWR models. We trained and evaluated the proposed models against sentence-based and word-based datasets. We utilized different strategies and structures to improve the performance of models. Our transformer-based models are actual E2E models following the S2S architecture with a self-attention encoder (SAE) and two decoders, a self-attention decoder (SAD), and a CTC decoder. Despite we did not incorporate any external modules such as an LM nor a dictionary into our architecture design, the proposed models are capable of recognizing complete sentences and words. Compared to state-of-the-art models, our transformer models have outperformed RNN models, which do not use LMs. Our best SAE-SAD model achieved a 5% CER and 18% WER against the CHAW dataset and 22% CER and 56% WER against the Online-KHATT dataset. Planned future work will involve investigating other features and expanding datasets by synthesizing new samples. We also plan to incorporate LM with transformer-based models to boost the performance.

REFERENCES

- [1] F. Alwajih, E. Badr, and S. Abdou, "Writer adaptation for e2e arabic online handwriting recognition via adversarial multi task learning," *Egyptian Informatics Journal*, 2022.
- [2] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.
- [3] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath *et al.*, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal processing magazine*, vol. 29, no. 6, pp. 82–97, 2012.
- [4] X.-Y. Zhang, Y. Bengio, and C.-L. Liu, "Online and offline handwritten chinese character recognition: A comprehensive study and new benchmark," *Pattern Recognition*, vol. 61, pp. 348–360, 2017.
- [5] V. Carbune, P. Gonnet, T. Deselaers, H. A. Rowley, A. Daryin, M. Calvo, L.-L. Wang, D. Keysers, S. Feuz, and P. Gervais, "Fast multi-language lstm-based online handwriting recognition," *International Journal on Document Analysis and Recognition (IJ DAR)*, vol. 23, no. 2, pp. 89–102, 2020.
- [6] S. Tabassum, N. Abedin, M. M. Rahman, M. M. Rahman, M. T. Ahmed, R. Islam, and A. Ahmed, "An online cursive handwritten medical words recognition system for busy doctors in developing countries for ensuring efficient healthcare service delivery," *Scientific reports*, vol. 12, no. 1, pp. 1–13, 2022.
- [7] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in *Proceedings of the 23rd international conference on Machine learning*, 2006, pp. 369–376.
- [8] J. Zhang, J. Du, and L. Dai, "A gru-based encoder-decoder approach with attention for online handwritten mathematical expression recognition," in *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, vol. 1. IEEE, 2017, pp. 902–907.
- [9] A. Hannun, C. Case, J. Casper, B. Catanzaro, G. Diamos, E. Elsen, R. Prenger, S. Satheesh, S. Sengupta, A. Coates *et al.*, "Deep speech: Scaling up end-to-end speech recognition," *arXiv preprint arXiv:1412.5567*, 2014.

- [10] D. Amodei, S. Anantharayanan, R. Anubhai, J. Bai, E. Battenberg, C. Case, J. Casper, B. Catanzaro, Q. Cheng, G. Chen *et al.*, “Deep speech 2: End-to-end speech recognition in english and mandarin,” in *International conference on machine learning*. PMLR, 2016, pp. 173–182.
- [11] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, . Kaiser, and I. Polosukhin, “Attention is all you need,” *Advances in neural information processing systems*, vol. 30, 2017.
- [12] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, “An image is worth 16x16 words: Transformers for image recognition at scale,” *arXiv preprint arXiv:2010.11929*, 2020.
- [13] A. Mohamed, D. Okhonko, and L. Zettlemoyer, “Transformers with convolutional context for asr,” *arXiv preprint arXiv:1904.11660*, 2019.
- [14] S. A. Mahmoud, H. Luqman, B. M. Al-Helali, G. BinMakhashen, and M. T. Parvez, “Online-khatt: an open-vocabulary database for arabic online-text processing,” *The Open Cybernetics & Systemics Journal*, vol. 12, no. 1, 2018.
- [15] O. Khaled, A. Fahmy, and S. Abdou, “Large vocabulary hybrid dnn/hmm arabic online handwriting recognition system,” in *2017 4th IAPR Asian Conference on Pattern Recognition (ACPR)*. IEEE, 2017, pp. 876–881.
- [16] J. Hu, M. K. Brown, and W. Turin, “Hmm based online handwriting recognition,” *IEEE Transactions on pattern analysis and machine intelligence*, vol. 18, no. 10, pp. 1039–1045, 1996.
- [17] M. Liwicki and H. Bunke, “Hmm-based on-line recognition of handwritten whiteboard notes,” in *Tenth international workshop on frontiers in handwriting recognition*. Suvisoft, 2006.
- [18] Y. Bengio, Y. LeCun, C. Nohl, and C. Burges, “Lerec: A nn/hmm hybrid for on-line handwriting recognition,” *Neural computation*, vol. 7, no. 6, pp. 1289–1303, 1995.
- [19] M. Schenkel, I. Guyon, and D. Henderson, “On-line cursive script recognition using time-delay neural networks and hidden markov models,” *Machine Vision and Applications*, vol. 8, no. 4, pp. 215–223, 1995.
- [20] M. Liwicki, A. Graves, S. Fernández, H. Bunke, and J. Schmidhuber, “A novel approach to on-line handwriting recognition based on bidirectional long short-term memory networks,” in *Proceedings of the 9th International Conference on Document Analysis and Recognition, ICDAR 2007*, 2007.
- [21] W. Yang, L. Jin, Z. Xie, and Z. Feng, “Improved deep convolutional neural network for online handwritten chinese character recognition using domain-specific knowledge,” in *2015 13th international conference on document analysis and recognition (ICDAR)*. IEEE, 2015, pp. 551–555.
- [22] P. S. Mukherjee, B. Chakraborty, U. Bhattacharya, and S. K. Parui, “A hybrid model for end to end online handwriting recognition,” in *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, vol. 1. IEEE, 2017, pp. 658–663.
- [23] J. Wang, J. Du, J. Zhang, B. Wang, and B. Ren, “Stroke constrained attention network for online handwritten mathematical expression recognition,” *Pattern Recognition*, vol. 119, p. 108047, 2021.
- [24] M.-T. Luong, H. Pham, and C. D. Manning, “Effective approaches to attention-based neural machine translation,” *arXiv preprint arXiv:1508.04025*, 2015.
- [25] A. D. Le, H. T. Nguyen, and M. Nakagawa, “Recognizing unconstrained vietnamese handwriting by attention based encoder decoder model,” in *2018 International Conference on Advanced Computing and Applications (ACOMP)*. IEEE, 2018, pp. 83–87.
- [26] J.-Y. Ye, Y.-M. Zhang, Q. Yang, and C.-L. Liu, “Contextual stroke classification in online handwritten documents with graph attention networks,” in *2019 International Conference on Document Analysis and Recognition (ICDAR)*. IEEE, 2019, pp. 993–998.
- [27] A. Al-Salman and H. Alyahya, “Arabic online handwriting recognition: a survey,” in *proceedings of the 1st international conference on internet of things and machine learning*, 2017, pp. 1–4.
- [28] B. M. Al-Helali and S. A. Mahmoud, “Arabic online handwriting recognition (aohr) a survey,” *ACM Computing Surveys (CSUR)*, vol. 50, no. 3, pp. 1–35, 2017.
- [29] F. Biadsy, J. El-Sana, and N. Y. Habash, “Online arabic handwriting recognition using hidden markov models,” 2006.
- [30] H. Ahmed and S. A. Azeem, “On-line arabic handwriting recognition system based on hmm,” in *2011 International Conference on Document Analysis and Recognition*. IEEE, 2011, pp. 1324–1328.
- [31] H. A. Abd Alshafy and M. E. Mustafa, “Hmm based approach for online arabic handwriting recognition,” in *2014 14th International Conference on Intelligent Systems Design and Applications*. IEEE, 2014, pp. 211–215.
- [32] I. Hosny, S. Abdou, and A. Fahmy, “Using advanced hidden markov models for online arabic handwriting recognition,” in *The First Asian Conference on Pattern Recognition*. IEEE, 2011, pp. 565–569.
- [33] M. Kherallah, N. Tagougui, A. M. Alimi, H. El Abed, and V. Margner, “Online arabic handwriting recognition competition,” in *2011 International Conference on Document Analysis and Recognition*. IEEE, 2011, pp. 1454–1458.
- [34] N. Tagougui, H. Boubaker, M. Kherallah, and A. M. Alimi, “A hybrid nn/hmm modeling technique for online arabic handwriting recognition,” *arXiv preprint arXiv:1401.0486*, 2014.
- [35] F. Alwajih, E. Badr, S. Abdou, and A. Fahmy, “Deeponkhatt: An end-to-end arabic online handwriting recognition system,” *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 35, no. 11, p. 2153006, 2021.
- [36] R. Maalej, N. Tagougui, and M. Kherallah, “Online arabic handwriting recognition with dropout applied in deep recurrent neural networks,” in *2016 12th IAPR Workshop on Document Analysis Systems (DAS)*. IEEE, 2016, pp. 417–421.
- [37] R. Maalej and M. Kherallah, “Improving the dblstm for on-line arabic handwriting recognition,” *Multimedia Tools and Applications*, vol. 79, no. 25, pp. 17969–17990, 2020.
- [38] S. Jaeger, S. Manke, J. Reichert, and A. Waibel, “Online handwriting recognition: the npen++ recognizer,” *International Journal on Document Analysis and Recognition*, vol. 3, no. 3, pp. 169–180, 2001.
- [39] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, “Distributed representations of words and phrases and their compositionality,” *Advances in neural information processing systems*, vol. 26, 2013.
- [40] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” *arXiv preprint arXiv:1409.1556*, 2014.
- [41] J. Gehring, M. Auli, D. Grangier, D. Yarats, and Y. N. Dauphin, “Convolutional sequence to sequence learning,” in *International Conference on Machine Learning*. PMLR, 2017, pp. 1243–1252.
- [42] 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.
- [43] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, “Dropout: a simple way to prevent neural networks from overfitting,” *The journal of machine learning research*, vol. 15, no. 1, pp. 1929–1958, 2014.
- [44] P. Izmailov, D. Podoprikin, T. Garipov, D. Vetrov, and A. G. Wilson, “Averaging weights leads to wider optima and better generalization,” *arXiv preprint arXiv:1803.05407*, 2018.
- [45] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin *et al.*, “Tensorflow: Large-scale machine learning on heterogeneous distributed systems,” *arXiv preprint arXiv:1603.04467*, 2016.
- [46] D. Wilson-Nunn, T. Lyons, A. Papavasiliou, and H. Ni, “A path signature approach to online arabic handwriting recognition,” in *2018 IEEE 2nd International Workshop on Arabic and Derived Script Analysis and Recognition (ASAR)*. IEEE, 2018, pp. 135–139.
- [47] Y. Hamdi, H. Boubaker, and A. M. Alimi, “Data augmentation using geometric, frequency, and beta modeling approaches for improving multi-lingual online handwriting recognition,” *International Journal on Document Analysis and Recognition (IJAR)*, vol. 24, no. 3, pp. 283–298, 2021.
- [48] Y. Hamdi, H. Boubaker, B. Rabhi, W. Ouarda, and A. Alimi, “Hybrid architecture based on rnn-svm for multilingual online handwriting recognition using beta-elliptic and cnn models,” 2021.

Detection of Android Malware App through Feature Extraction and Classification of Android Image

Mohd Abdul Rahim Khan^{1,2}

Department of Computer Science¹,

College of Computer and Information Sciences,

Majmaah University, Majmaah, 11952, Saudi Arabia

Department of Computer Science and Engineering²,

Lingaya's Vidyapeeth, Faridabad,

Haryana, India

Prof. Nand Kumar

Department of Computer Science
and Engineering,

Lingaya's Vidyapeeth, Faridabad,

Haryana India

R C Tripathi

Department of Computer Science
and Engineering, Teerthanker Mahaveer

University, Moradabad

Abstract—Android apps have security risks due to rapid development in android devices. In the Android ecosystem, there are many challenges to detecting Android malware. Traditional techniques such as static, dynamic, and hybrid approach, most of the existing approaches require a high rate of human intervention to detect Android malware. Most of the current techniques have the most significant security challenges to detect Android malware, the inspection of Android Package Kit(APK) file structures, increased complexity, high processing power, more storage space, and much human intervention. This paper proposed Machine Learning(ML)based algorithms to detect Android malware apps through feature extraction and classification of grayscale images. In our proposed approach, convert most of the files of APK such multiDex, resources, certificate, and manifest transform into a grayscale image, using the image algorithm to extract the local feature of the image. In the paper used different ML models to classify the local features with the help of multiple images of malware families. This approach deals with the obfuscation attack. It can hide in any files of APK. The proposed approach enhanced accuracy reached up to 96.86%, and computation time did not increase more than the existing techniques. The quality of that proposed worked; it has a high classification accuracy and less complexity validation loss.

Keywords—Android malware; obfuscation attack machine learning; android application package (APK); android malware app; grayscale images

I. INTRODUCTION

Android operating system (OS) is the most popular OS in the smart device ecosystem. Due to intelligence devices, every android user is very close to and dependent on Android Application Package (APK). In the present scenario, the android users sharing sensitive information, banking operation, e-shopping, locations information, the identity of the users, and privacy of data are also involved. In Android, device security is the biggest challenge and severe issue. A survey report of GDATA in 2019 [1] showed that 1,852,170 Android malware samples were detected in the first half of 2019. Here, data showed an android malware is detected every 8 seconds. The statistical report represents eight mobile infected by malicious out of ten Android devices [2]. One more research report, Google detected 86% of the total Android devices market in 2017. The most popular GlobalStats website showed that 73% of Android-based devices counted sales of total devices of Android in 2019 [3]. Due to the popularity of Android

devices, Android app becomes more targeting apps compared to other kinds of apps. As per one evolution report of mobile malware, 5,321,142 apps were installed on devices, 151,359 mobile apps were detected as Trojans, 60,176 were detected as mobile ransomware by Kaspersky 2018 [4].

Android users threaten by different types of malware families; some are distributed by Google Play stores, some type apps such as downloader, banker, and hidden ads [5]. Most of the extensive attacks pointing the Android OS. The hackers mainly focused on attacking games, banking, academics, e-shopping domain. However, this domain published many malicious apps, which have gaps between the app development and the number of works. Third-party stores have untrusted apps; most gaming apps have adware due to the repackaging technique [6]; with the help of repackaging tools, reassemble the original app and add the malicious code with the original code and then assemble again, upload on third party store. Here, the main challenging task to identify malicious apps is the most severe issue. Most existing techniques are static and dynamic; most techniques used behavior and signature base to identify the android malware.

Static techniques do not require running apps; they disassemble the code and extract the feature of apps to identify. The dynamic approach always needs to run the application and identify the android malware through behavior and signature base. These techniques have significant drawbacks; it requires more computing power, resources, and space [7]. The dynamic analysis was evaded by some powerful and intelligent malware [8]. Moreover, existing dynamic and static techniques used the manual intervention of humans. It also needs domain expertise to identify reverse engineering [9]. The existing approach used single classes.dex file, but in the current scenario, we have multiclass files, or multiDex files [10], which have not been converted into a grayscale image to detect malware.

Our proposed work takes care of all essential files such as multiDex (MD), resources.ARSC(RS), Manifest.xml(MX), and certificate (CR) files of APK to detect Android malware. The existing approaches need human intervention to separate the dex, manifest.xml resource.ARSC(RS) and certificate files convert into the grayscale image [11-12].

META-INF: This file very essential in Android apps, which the information about the signature and information

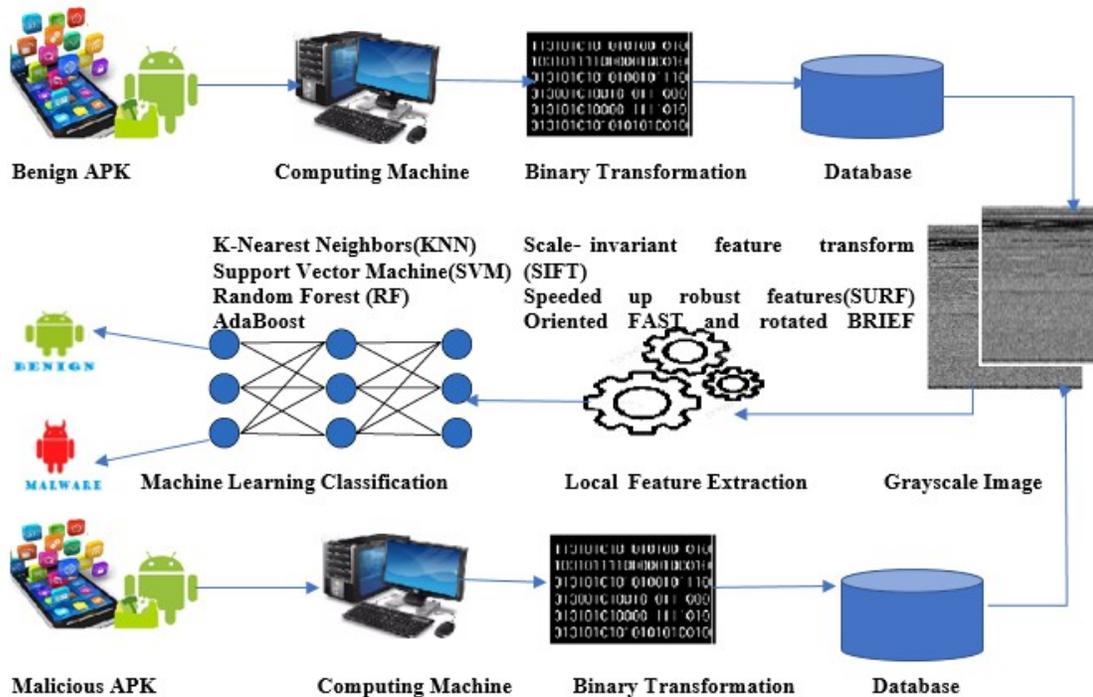


Fig. 1. Proposed Methodology to Detect the Android Malware

about the resource list.

Lib: lib file is used to run the specific device architectures of the native library, such as armeabi-v7a and x86.

Res: to Keep the resources such as images. Which is not compiled with resources.arsc

Assets: Raw information about resources

AndroidManifest.xml: Meta information about the apps such as version, content, and name of APK files.

Multiple classes.dex: Main and necessary file of apps, which run java class methods on the devices.

Resources.arsc: Compiled all resources on the devices which is used by the apps.

Android apps development using java.class files. By the DX tools convert multiple java.class into the DEX files. DEX and manifest are essential files in APK, and DEX consists of the data structure; the interpreter used the different data types that belong to the data structure. All static reverse engineering tools used the DEX files to reassemble the apps for reengineering. Multiple methods are proposed to protect the DEX files. Our Proposed approach does not require any human intervention, does not require separate files, and does not need reverse engineering to find different types of files. Our proposed approach takes less computation power to detect the android malware because it takes less time complexity because it worked without any reverse engineering operation. Our proposed approach used DEBIAN and AMD datasets containing 10560 apps (5000 benign and 5560 malicious apps). The grayscale image datasets, each containing 10560 samples (5000, 5560 benign, malicious samples, respectively), were constructed based on diverse files from the contents of the APK

collections. Firstly, all the benign and malicious APK convert into Grayscale images, a block diagram depicted in Fig. 2. Secondly, extract the local features from images using image-based feature extraction techniques such as SIFT, SURF, and ORB. Thirdly, apply the BOVW approach to convert multiple local feature descriptor vectors into a single feature vector to feed into ML classifiers. Finally, extract the global and local features and apply the different ML classifier techniques such as AdaBoost K-Nearest Neighbors(KNN), Support Vector Machine(SVM), and Random Forest(RF). The Proposed approach worked on the raw bytes of grayscale images; the main advantage of this approach does not require any reengineering operation and making different types of datasets. The existing approaches have the main disadvantage, approaches that require human intervention. Our approach proposed safe from human intervention and reengineering operation. Many ML algorithms are developed for the detection of malware apps. The most common challenge in Android malware detection is obfuscation attacks. Malicious code can be hidden in any files of APK, which is very dangerous to android malware app detection. Our proposed works have a novelty that now no needs to do reverse engineering to obtain all files of APK. Directly conversion of Entire APK files structure converts into a grayscale image. Most of the existing techniques used separate files to transform into grayscale scale images to analyze the image-based android malware detection. All existing methods do not care about multiple DEX, Share Object (SO), Meta-Inf, lib files, etc., just observation of manifest, single DEX files, resource files only. In the meantime, the author should explain the functions of multiDEX (MD). Resources, ARSC (RS), Manifest.xml (MX), and certificate (CR) files of APK separately because they are used to detect Android malware. Then, the proposed methodology to detect the android malware

is well represented in Fig. 1.

II. RELATED WORK

Many researchers worked in the domain of Android malware detection; some are listed below in this section. An approach designed to analyze the suspicious behaviors and detection of resources abuse [13]. The major drawback of this approach is the need to decompile the app and embedded hook code; this approach used runtime events to track and monitor the logging. The SafeDroid static framework approached, which statically analyzed the DEX (Dalvik Executable). By this approach, extract the binary feature vectors to train various ML classifiers [14]. Moreover, the multiple features are system calls, app permission, system events [15]. Those features train RF classifiers to analyze whether Apps are affected by malicious or not. Some approaches differentiate whether the app is a malicious or normal app based on patterns permission [16], the required permissions extracted statically. Most popular permissions are registered into class [17] to define whether the permission is benign or malicious. The permissions of a class determine the benign and malicious app. Moreover, a data mining technique made the constructive pattern of permission to determine whether the android app is malicious or benign. Here, the authors applied the bi-clustering method to used permissions. Also, the authors used the information of the Android app package and permissions to train of KNN, Linear Discrimination(LD) function, and Radial Basis Function (RBF) network. Moreover, Application Programming Interface(API) system calls integrated with permissions [18] are used as features to train the RF classifier of android's apps classification. It is a very lightweight method for detecting Android malware through ML and dataflow-related API system calls used in this approach [19]. In [20], the proposed approach used the n-gram series to extract the features from the opcode of malicious and benign apps. This approach used a limited number of features to train RF and Support Vector Machines (SVM) classifiers. The proposed approach [21] installed the Android application(APK) on Android devices to extract dynamic features such as networks behavior, memory consumption, computation power, time-space, battery, and binder; these features are used to classify malware. This dynamic approach [22] captured network traffic behaviors of running Android applications(APK) from different android devices. This traffic correlates with malware URLs and with DNS service network traffic for the detection of malware. An approach [23] used to fog computing reduces the load and dynamically enhances the computation power to detect Android malware. Another approach [24] used the API system calls and network behaviors, collectively applied to detect Android malware. In [25] this paper, the authors showed the multiple network behavior and emulator-based dynamic experiments to analyze android malware. Android operating system embedded by an extension kit has been proposed [26] to deals with confused delegate attacks [a genuine APK is manipulated for communicating with the trusted application for Inter-Process Communication(IPC)]. To enhance permissions-based policy [27]at runtime tracking and communication link analysis by pre-defined policy to prevent malicious behavior. Moreover, the signature set is constructed by network log and correlated with the permissions-based methods [28] for android apps classification. The recent approach [29] uses

reverse engineering techniques to decompile the APK, extract the source code, and convert it into a grayscale image. The constructed dataset of images is used to train a convolutional neural network (CNN) to detect the malicious app. API system calls and semantic information is used to train the Short-Term Long Memory (LSTM)[30] model to classify the android malware. Moreover, a hybrid approach includes CNN and deep autoencoder (DAE) [31] to detect Android malware. In [32], this proposed approach used the hybrid scheme; it extracts dynamic and static behavior features used to train the deep learning model. Also, in [33] approach extracted the four features, such as permissions, rate of permission, system events, APIs system calls used to train the collective RF classifier. DREBIN [34] is a static analysis approach; this approach used similar malicious apps as per experiment works (5,560 malicious apps). This method used as many possible features of apps and was added with joint vector space. Due maximum number of features and determination increased the complexity level. This paper [35] proposed the classification of the dependency graph. The features extracted from the dependency graph make the semantic feature set from the weighted contextual API of the graph. The metric of the homogeneous app determines same the application behaviors. The sensitive and important API call allocated the weight according to the Android malware family [36]. Every app implemented the function call graph (FCG), and each FCG construct the sensitive API call-related graph (SARG). The SARG has the parent and sensitive API call nodes. Here, train multiple machine learning approaches to classify the common behavior of the malware family. Moreover, from source code is extracted from hexadecimal representation and converted into RGB images [37]. The color RGB dataset is used to train a CNN classifier to classify Android malware. Furthermore, the Android (APK) application converted to grayscale images, then extract the feature of grayscale images for training the RF classifier for classification in [38]. Also, in [39], extracted the feature from 2D of Opcode Sequences and assigned the weight based on their occurrence. The weight value is converted into grayscale images. The image detection approach is very limited to detecting Android malware domains. Local and global feature extraction of the entire APK is more effective than existing approaches. Our paper has mainly converted the image into grayscale without any reverse engineering tools. It does not require separating the files of APK such as resource, Multidex, manifest, and certificate. Moreover, it does not require human intervention; most existing techniques have a common issue of human intervention and extracting the source code from reverse engineering tools.

III. METHODOLOGY

This section discussed the full detail of the proposed model. The first subsection briefly describes constructed dataset, the other section described the brief details of extracted features, and the last section briefly describes the training Machine Learning (ML) classification.

1) *Dataset*: this dataset. Our experiment setup used the 5560 files of android malware and 5000 benign apps from AMD, which have 179 different Android malware families. In the investigation of research of android malware from 2012 to 2020, most of the researchers used the DREBIN dataset. The DREBIN dataset has the most famous malware such as

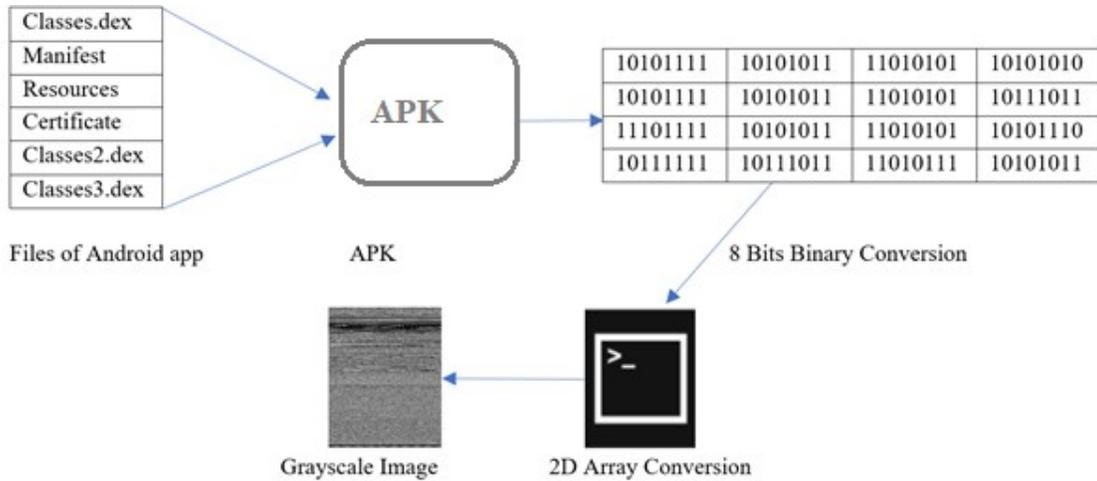


Fig. 2. APK to Image Conversion Process

DroidKungFu, GingerMaster, GoldDream, and Fake Installer. The primary objective of our proposed approach is to detect Android malware. The DREBIN dataset has the most famous malware such as DroidKungFu, GingerMaster, GoldDream, and Fake Installer. The primary objective of our proposed approach is to detect Android malware. Many researchers used the DREBIN dataset to analyze the android malware, and various institutions utilized this dataset to investigate Android malware. Our Proposed models used the DREBIN dataset because it has 179 different android malware families, appropriate for any investigation dataset.

A. Transformation APK into Grayscale Images

The Android APK files convert into grayscale images [40]. In this proposed article, the authors construct the malware images using files of the Android app from malware APK. The APK is transformed into 8-bit vectors, and then the 8-bit vector transforms into a grayscale image. Every substring has an 8-bit value as a pixel converted into a decimal value between 0-255, shown in Fig. 2. Any digital file on the memory device is stored as a stream of a bit of '0' and '1'. In the model read every APK file as a binary stream, group every eight bits, and store them in a new file with the image file extension.

B. Local Features Extraction

The local feature is a defined image object (basically, in the image, a cluster of pixels or small blobs) [41]. The local feature of images is the most stimulating point in the image, which defines the image descriptor vectors(DV) or feature vectors. The set of feature vectors is described by different types of algorithms. Our proposed approach used the four different algorithm types to extract the local features as Scale invariant feature transform(SIFT), Speeded up robust features (SURF), Oriented FAST, and Rotated BRIEF(ORB). Those methods are very famous in the malware domain for better accuracy.

1) *SIFT*: The Scale invariant feature transform (SIFT) method is applied to extract the local feature key points. This method computes the Laplacian of Gaussian on the multiple

scale level to provide a better result. This SIFT algorithm obtains the local minima and local maxima of stimulating points with the help of LoG at different scale levels. The several Laplacian of Gaussian on different scale levels (ρ), by the scale, obtain the local maxima, minima of every single pixel in the image. The Laplacian of Gaussian (LoG) calculation is costly for feature points to more or less extent. The Laplacian of Gaussian(LoG) approximately is determined by Eq.1.

$$\rho \nabla^2 L = \frac{\partial L}{\partial \rho} = \frac{L(x, y, k\rho) - L(x, y, \rho)}{k\rho - \rho} \quad (1)$$

where $L(x,y,\rho)$ is the Laplacian Gaussian on the position (x, y) at scale ρ . $L(x,y,k\rho)$ is the Laplacian of Gaussian(LoG) on the position (x, y) at scale $k\rho$, and the $k\rho$ is a scale a little more than ρ . The SIFT methods identified the stimulating points at the level of 128-bit descriptors in Eq.1. The extracted feature from the input images through the SIFT matched each feature of k nearest neighbors. The main objective of SIFT is to object recognition techniques to panorama stitching. As a result, the system is insensitive to the images' ordering, positioning, scale, and illumination. Two-Dimension isotropic measure by the Laplacian to the second spatial derivative of an image. The Laplacian Gaussian approach highlights areas of speedy intensity change and is often used for zero-crossing edge detectors. In our system, the Gaussian smoothing filter reduces its sensitivity to noise for smoothing with something approximating.

2) *SURF*: The algorithm that Speeds up robust features(SURF) [42] is the faster algorithm, and it can be the replacement for SIFT. This algorithm is faster and more robust for similarity comparison and similarity invariant of images. SURF algorithm plays a vital role in the real type of tracking and recognition of the object. The main merit of this algorithm is box filters approximation and calculation of the integral images. Additionally, it has the location and scale-based determinant of the Hessian matrix. The Hessian matrix has good performance to obtain the image key points, and it has good accuracy. In the SURF algorithm filtered by Gaussian

kernel, with location $X=(x,y)$, and scale ρ in Eq.2.

$$H(x, y) = \begin{pmatrix} S_{xx}(x, \rho) & S_{xy}(x, \rho) \\ S_{xy}(x, \rho) & S_{yy}(x, \rho) \end{pmatrix} \quad (2)$$

where, $S_{xx}(x, \rho)$ has a Gaussian kernel derivative on the point of x in the image, and similarly for $S_{xy}(x, \rho)$ and $S_{yy}(x, \rho)$. Haar-wavelet responses determine horizontal and vertical paths to the neighborhood of size six and used the 64 Bit Descriptor. Within interest point neighborhood, distribution of Haar-wavelet responses obtained from descriptor description. We need integral images to speed up the system. Additionally, using the 64 Bit Descriptor dimensions to improve the system's performance for feature computation increases robustness and matching. In the invariant to rotation, we recognized the reproducible orientation for the interest points. For this reason, we obtained the Haar-wavelet responses in the vertical and horizontal directions. The circular neighborhood of radius 6s around the interest points, with s the scale that the interesting point detected. Therefore, our proposed approach uses integral images for fast filtering again. Only six actions are needed to SURF: Speeded Up Robust Features, the seventh determines the feedback in the vertical and horizontal directions at any scale.

3) *ORB*: The feature vector Oriented FAST and rotated BRIEF (ORB) is a high-speed keypoint detector [43]; in BRIEF, descriptors have much modification to improve the algorithm performance. The ORB algorithm detects the keypoint in images by using the FAST algorithm. Also used the Harris corner to detect the key point. Moreover, it used the multiscale feature with 32 bits BRIEF-based descriptor

$$(S;x, y) = \begin{cases} 1 : S(x) < S(y) \\ 0 : S(x) \geq S(y) \end{cases} \quad (3)$$

where S is the flattened spot in the image, and $S(x)$ is the intensity in Eq. 3. In the implementation of the FAST algorithm, we extract the kernel windows from single line buffers. In the approach, the center pixel is subtracted from each circle pixels. The result is measured with the minContrast value whenever the obligatory number of consecutive pixels exceeds the threshold level; the center is marked as the corner. For the circle region, evaluate the sum-of-absolute-difference (SAD) metric. Only the differences that exceed the minimum contrast threshold level are involved in the metric. This calculation means that the algorithm detects a light center pixel surrounded by dark pixels or a dark center pixel surrounded by light pixels as corners with high metrics. The Harris algorithm used five image filters, and three circular windows and evaluated the two gradients. The design of the calculation of the eigenvalue of the Harris matrix practices three multipliers and three adders and is pipelined to optimize performance.

C. Machine Learning (ML) Classification

Our proposed models used four types of Machine Learning models such as Adaboost, K-Nearest Neighbors(KNN), Support Vector Machine(SVM), and Random Forest(RF) to classify the extracted local features from Grayscale images.

1) *K-Nearest Neighbors (KNN)*: K-Nearest Neighbors(KNN) is a supervised ML models, which is used for the classification of input data. It recognizes data points classified into multiple classes and calculates the class label for the new input data point. This method is famous for classifying the object into the train closest feature space. The nearest neighbors are signified by K in KNN, and the maximum unknown data points classify near to K neighbors. The primary benefit of the KNN algorithm uses the minimum distance to search the nearest neighbors. The selection of the number of nearest neighbors is essential to obtain the augmented KNN model. The selection of the number of nearest neighbors is essential to get the augmented KNN model.

2) *Support Vector Machine (SVM)*: Support Vector Machine(SVM) also is a supervised ML algorithm. In this model, take the past input data and predict the feature output. The primary purpose of SVM is classification, but it is also used for regression statements. The SVM algorithm chooses the support vectors in the dataset at the extreme points. It selects the maximum distance between the support vector and hyperplane as much as possible. A class in support vectors has the maximum distance from the hyperplane. The distance margin defines as the distance between different support vector classes. The sum of $D+$ and $D-$ is calculated as distance margin, where $D-$, hyperplane has the minimum distance from the closest negative point and $D+$, hyperplane has the minimum distance from the closest positive point. The main aim of SVM is to find the maximum distance margin, which gives the optimal hyperplane. The optimal hyperplane always gives excellent classification. In the case of non-linear, which produces low and no distance margin, SVM showed misclassification. In that scenario, SVM used the kernel functions to convert the non-linear data into 2D or 3 D dimension arrays. The minor dimensional feature is converted into high dimensional feature space by the kernel functions.

3) *Random Forest (RF)*: Random Forest (RF) is one of the most common and powerful supervised ML algorithms. RF executes efficiently massive datasets and predicts accurate results. This algorithm support both types of functionality, such as classification and regression—the decision tree support RF to enhance the accuracy and flexibility. In general, with more trees in the forest, the output would be more predictable. The more trees in the RF reduce the risk of when a statistical model fits exactly against its training data. RF can obtain good accuracy in case of missing a large proportion. According to attributes, the new object classifies, and the decision tree gives the classification output per the ruleset.

4) *AdaBoost*: The first boosting algorithm is AdaBoost, which solved multiple problems. The AdaBoost constructs a robust classifier from multiple weak classifiers. This algorithm keeps a single split of the decisions tree with the weak stump, known as the decision stump. AdaBoost always keeps more load on tough to classify, easy to handle the problem, and do less. This algorithm has solved both types of problems, such as classification and regression. Multiple APK's are repackaged, which steal code by reverse engineering methods and reassemble with another name by adding adware or small scripts of malicious code into repacked APK. Here the APK has very slightly changed, so the dataset has slight noise in

data We found that in the case of less noisy data, only a few hyperparameters need to be tuned to improve the Adaboost performance. In the case of the small number of input variables, KNN models provide excellent performance. Whenever we increase more number of input variables, the performance of KNN degrades. In our dataset, we used multiple DEX files based on APKs. All files structure of APK were converted into grayscale images, which increased the number of input variables and memory size and complexity of the KNN model.

IV. PROPOSED MODELS

In our work, we proposed an image-based detection of android malware using machine learning classification. In this process, Android APK converts into a grayscale image, extracts the image feature using image processing techniques, and trains the machine learning classification to detect malicious or benign apps depicted in Fig. 1. The novelty of this approach the entire files of APK transforms into images to deal with the obfuscation attack. Most of the existing techniques used only three files of APK to transform into the image. The main disadvantage of the other techniques requires decompiling the APK and separating the files such as DEX (MD), ARSC(RS), Manifest.xml(MX), and certificate files. Moreover, the disadvantage is that it does not take care of the mutliDEX files. If APK has more than 6500 methods in the app, it needs to create the multiDex files [40]. If the malicious code is embedded with second or third classes.dex files, no existing algorithm detects the Android malware app from multiDex class files. The primary source of the malicious code is embedded into classes.dex files. Our models used three algorithms (SURF, ORB, and SIFT) to extract local features (LF) descriptors from the grayscale image dataset. One by one, local features (Extracted from each image) train to multiple machine learning algorithms (RF, KNN, DT, and AdaBoost). The multiple descriptors represent an image. Above mentioned machine learning algorithm gave the multiple vectors as outputs, which cannot be direct as inputs for any machine learning algorithm. This model used the Bag of Visual Words(BOVW) to create one feature vector with multiple local feature descriptors [41]. The BOVW uses any clustering techniques to fragment the extracted descriptors vectors into multiple clusters. Then the cluster is predicted by the clustering algorithm.

A. Accuracy Assessment

The accuracy metrics for multiple Machine Learning models were determined based on the precision, recall, f1-score, and accuracy in Eq. 4, 5, 6, 7 respectively, and precision in fraction of data entries of malicious activity are categorized as truly Android malware.

$$\text{Precision} = \frac{\text{True Possitive}}{\text{True Possitive} + \text{False Possitive}} \quad (4)$$

The recall is the fraction of malicious apps data of correctly classified malicious families.

$$\text{Recall} = \frac{\text{True Possitive}}{\text{True Possitive} + \text{False Negative}} \quad (5)$$

TABLE I. OBTAINED ACCURACY, PRECISION, RECALL AND F1-SCORE FROM MULTIPLE MACHINE LEARNING MODELS USING MULTIPLE LOCAL FEATURES EXTRACTOR METHODS

Performance Evaluator	Feature Vectors Methods	Machine Learning Methods			
		K-Nearest Neighbors	Support Vector Machine	Random Forest	AdaBoost
Accuracy	SIFT	92.42%	94.06%	94.65%	93.16%
	SURF	94.69%	95.37%	96.33%	96.86%
	ORB	89.41%	89.83%	91.42%	92.83%
Precision	SIFT	94.48%	91.00%	95.11%	95.71%
	SURF	95.79%	94.63%	97.44%	97.41%
	ORB	90.46%	89.13%	92.48%	93.36%
Recall	SIFT	91.31%	94.97%	93.42%	92.66%
	SURF	93.47%	96.29%	95.09%	96.35%
	ORB	88.33%	90.70%	90.24%	92.34%
F1-Score	SIFT	92.40%	94.14%	94.57%	93.18%
	SURF	94.71%	95.45%	96.25%	96.88%
	ORB	89.39%	89.91%	91.34%	92.85%

f1- score is the harmonic mean between sensitivity and precision.

$$f1 - score = 2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \quad (6)$$

The accuracy or complete classification accuracy is the portion of all suitably classified negative and positive records with the losses.

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

$$\text{Cross-EntropyLoss} = -\frac{1}{N} \sum_{o=1}^n \log p_{model}(y_o \in C y_o) \quad (8)$$

V. EXPERIMENT

The experiment has been designed on Intel core™ i-7 10700 CPU @ 3.8 GHz with 16 GB RAM. The experiment used the BOVW algorithm, which needs a size 120 codewords vocabulary. The K-means technique collected all key points from created datasets, and it has the codeword vocabulary size 120. Moreover, the proposed model used Opencv, Sklearn python libraries for the implementation of laboratory works.

The performance of different Machine Learning models has been achieved in terms of the whole percentage of true positive, true negative, false positive, and false-negative decisions. Our local features extractor models, such as SIFT, SURF, and ORB, extract key points from the image dataset. The extracted local feature passed to train four renowned machine learning models, i.e., K-Nearest Neighbors(KNN), Support Vector Machine (SVM), Random forest, and AdaBoost. The complete result with multiple ML models and local feature extractor models is presented in Table I and shown in Fig. 3. The validation set of the accuracy and losses in our proposed works proves that the results are correct, not overfitting problems depicted in Fig. 4. The high accuracies, precision, recall, and Fi-score from different machine learning models are displayed in Table II and Fig. 3. if the losses decrement and accuracy developments of both groups are like the same, then the process aborted changed the modeling parameters to remove the overfitting problem. Last, the AdaBoost model accuracy touched 96.86%. The traditional machine-learning algorithm shows the performance of each algorithm in Table III and is

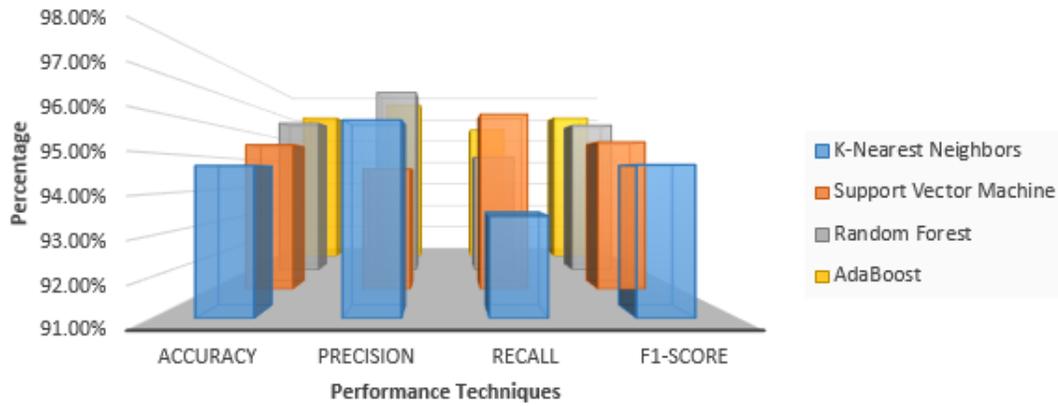


Fig. 3. Performance of Multiple Machine Learning Models

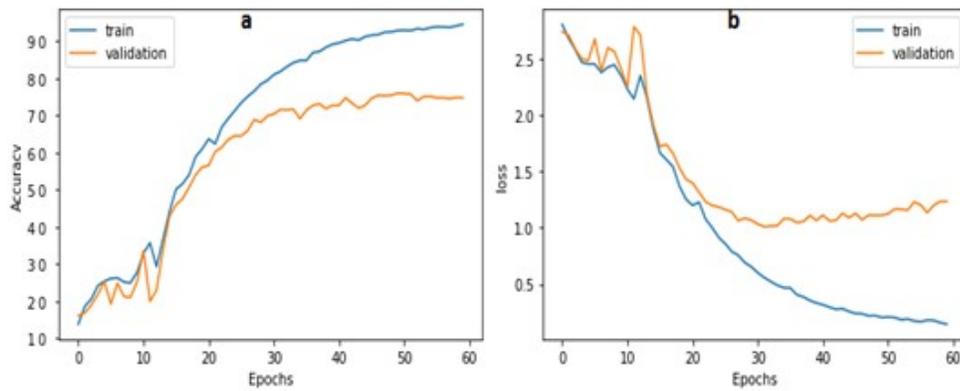


Fig. 4. (a): Train and Validation of AdaBoost Accuracy; (b): Train and Validation Loss of Model using SURF Local Feature

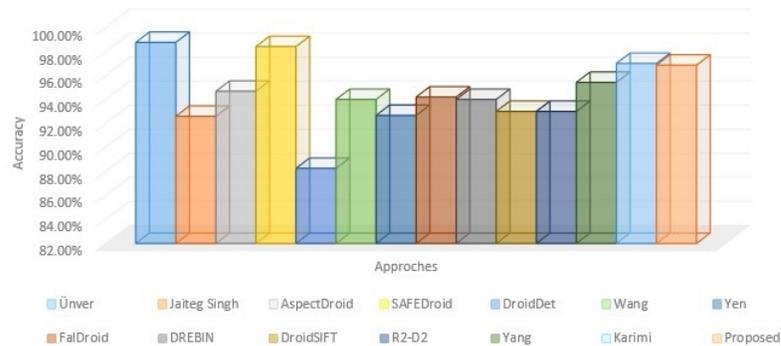


Fig. 5. Comparison of different Existing Approaches

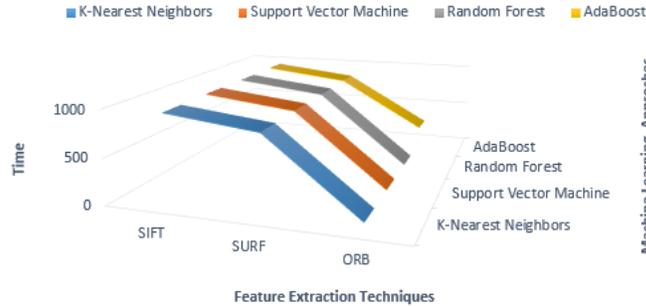


Fig. 6. Execution Time of the Proposed Model with Feature Extraction

TABLE II. PROPOSED MULTIPLE MACHINE LEARNING MODELS WITH ACCURACY, PRECISION, RECALL AND F1-SCORE

ML Method	Accuracy	Precision	Recall	F1-Score
K-Nearest Neighbors	94.69%	95.79%	93.47%	94.71%
Support Vector Machine	95.37%	94.63%	96.29%	95.45%
Random Forest	96.33%	97.44%	95.09%	96.25%
AdaBoost	96.86%	97.41%	96.35%	96.86%

TABLE IV. THE EXECUTION TIME OF THE PROPOSED MODEL

Feature Vectors Methods	Machine Learning Models			
	K-Nearest Neighbors	Support Vector Machine	Random Forest	AdaBoost
SIFT	941.63	945.87	941.03	943.91
SURF	827.00	830.64	826.72	828.91
ORB	43.36	44.87	44.10	266.43

TABLE III. COMPARISON OF DIFFERENT EXISTING APPROACHES

Model	Accuracy	Types
Unver [11]	98.75%	Image-based
Jaiteng Singh [12].	92.59%	Image-based
AspectDroid[13]	94.68%	Hybrid analysis
SAFEDroid[14]	98.40%	Static analysis
DroidDet[15]	88.26%	Static analysis
Wang[16]	94.00%	Static analysis
Yen[23]	92.67%	Image-based
FalDroid[33]	94.20%	Static analysis
DREBIN[34]	94.00%	Static analysis
DroidSIFT[35]	93.00%	Static analysis
R2-D2[36]	93.00%	Image-based
Yang[37]	95.42%	Image-based
Karimi[38]	97.00%	Image-based
Proposed	96.86%	Image-based

depicted in Fig. 4. The accuracy value maximum 96.88% in the AdaBoost model, with losses, using SURF local feature extraction models in Fig. 4. In our work, computation time is scanned in each step of all ML models, inclusive of the feature extraction process, training, and testing of the model. The computation time from different ML models, the feature extraction process, training, and validation are presented in Table IV and Fig. 6.

VI. CONCLUSION

Our proposed model uses an image-based framework to classify the android app, whether malicious or benign app. Most image-based detection techniques do not care about the multiDex files of APK, using only a single class.DEX file for image conversion. In existing techniques of image-based malware detection found [11-12,23,36,37] the maximum detection probability of malware in classes.DEX file, not in another file such as Resources.ARSC, Manifest.xml, certificate files, if the hackers hide malicious code into second or third classes.DEX files, there is no chance to detect the malware in previous approaches. In our experimental works, transform

all classes.DEX APK file's contents into grayscale images. We used the image processing techniques to extract the local feature of images, including SIFT, SURF, and ORB models. The Local features are classified using machine learning models (KNN, SVM, RF, and AdaBoost) to detect Android malware. The achieved results exhibited that the proposed approach overtakes the existing techniques in classification accuracy and computational time. Our work showed that the AdaBoost detection rate reached up to 96.86 %, shown in Fig. 5, and run time did not exceed 0.0195 s on average for each sample. In the future, we will try to use the local and global features of images on multiDEX files to classify the Android malware to improve accuracy.

ACKNOWLEDGMENT

Mohd Abdul Rahim Khan would like to thank the Deanship of Scientific Research at Majmaah University for supporting this work under Project No. R-2022-162

REFERENCES

- [1] G Data, Bochum, Germany, Tech. Rep. [Online] Access date 5/5/2021. Available: <https://www.gdatasoftware.com/mobile-internet-security-android>.
- [2] Gartner (2018) Gartner says worldwide sales of smartphones recorded first-ever decline during the fourth quarter of 2017.<https://www.gartner.com/en/newsroom/press-releases/2018-02-22-gartner-says-worldwide-sales-of-smart-phones-recorded-first-ever-decline-during-the-fourth-quarter-of-2017>. Accessed 27 Oct 2019.
- [3] StatcounterGlobalStats (2020) Mobile operating system market share worldwide. Mobile Operating System Market Share Worldwide <https://gs.statcounter.com/os-market-share/mobile/worldwide>. Accessed 09 Mar 2020.
- [4] SecureList (2018) Mobile malware evolution 2018. <https://securelist.com/mobile-malware-evolution-2018/89689/>. Accessed 27 Oct-2019.
- [5] DoctorWeb (2019) Doctor Web's overview of malware detected on mobile devices in September 2019." <https://news.drweb.com/show/review/?i=13446>. Accessed 27 Oct 2019.

- [6] M.R. Khan, R.C. Tripathi, and A. Kumar, Repacked android application detection using image similarity, *Nexo Revista Científica*, vol 33, no.1, pp.190-199, June, 2020.
- [7] W. Wang, M. Zhao, Z. Gao, Xu, G., H. Xian, Li, Y. and X. Zhang, Constructing features for detecting android malicious applications: issues, taxonomy and directions. *IEEE access*, vol.7, no. 2019, pp.67602-67631, June, 2019.
- [8] M.K.Alzaylaee, S.Y. Yerima, and S. Sezer, Emulator vs real phone: Android malware detection using machine learning. In *Proceedings of the 3rd ACM on International Workshop on Security and Privacy Analytics*, Scottsdale, Arizona USA , 2017, pp. 65-72.
- [9] B. Jung, T.Kim, and E.G Im, October. Malware classification using byte sequence information.. In *Proceedings of the 2018 Conference on Research in Adaptive and Convergent Systems*, Honolulu ,Hawaii ,2018, pp. 143-148.
- [10] M.R. Khan, and M.K Jain, Protection android app with multiDEX and SO files from reverse engineering, *Materials Today: Proceedings*, vol. 2021, no. 1, pp. 1-9, January ,2021.
- [11] H.M. Ünver, and K. Bakour, Android malware detection based on image-based features and machine learning techniques, *SN Applied Sciences*, vol 2, no.7, pp.1-15, June, 2020.
- [12] Singh, J., Thakur, D., Ali, F., Gera, T., & Kwak, K. S. Deep feature extraction and classification of android malware images. *Sensors*, vol 20, no. 24, p.7013, Jan, 2020.
- [13] A.I. Ali-Gombe, B. Saltaformaggio, D. Xu, and Richard III, G.G., Toward a more dependable hybrid analysis of android malware using aspect-oriented programming, *computers & security*, vol. 73, no. 2018, pp.235-248, Mar, 2018.
- [14] R. Goyal, A. Spognardi, N. Dragonian and M.Argyriou, SafeDroid: a distributed malware detection service for android. In *2016 IEEE 9th international conference on service-oriented computing and applications (SOCA)*, Macau, China 2016, pp. 59-66.
- [15] H. J . Zhu, Z. H . You, Z. X. Zhu, W. L. Shi, X Chen, & , L. Cheng , DroidDet: effective and robust detection of android malware using static analysis along with rotation forest model, *Neurocomputing*, vol 272,no. 2018, pp. 638-646, 2018.
- [16] C.Wang, Q. Xu, , X.Lin. , & S. Liu, Research on data mining of permissions mode for Android malware detection, *Cluster Computing*, vol. 22 , no. 6, pp. 13337-13350, Nov, 2019.
- [17] V. Moonsamy, J. Rong, S. Liu , Mining permission patterns for contrasting clean and malicious android applications, *Future Generation Computer Systems*, vol 36, pp. 122-132, Jul, 2014.
- [18] G.Tao, Z. Zheng, , Z. Guo, , & M. R. Lyu, MalPat: Mining patterns of malicious and benign Android apps via permission-related APIs, *IEEE Transactions on Reliability*, vol. 67, no. 1, pp. 355-369,Dec, 2017.
- [19] M. Turner, B.Kitchenham, , P. Brereton, , S. Charters, , & D. Budgen, Does the technology acceptance model predict actual use? A systematic literature review, *Information and software technology*, vol 52, no 5, pp. 463-479, May ,2010.
- [20] G.Canfora, A. De Lorenzo , E. Medvet, F. Mercaldo, & C. A. Visaggio, Effectiveness of opcode ngrams for detection of multi family android malware. In *2015 10th International Conference on Availability, Reliability and Security (IEEE)*, Toulouse, France, 2015, pp. 333-340.
- [21] H. Papadopoulos, , N. Georgiou, , C. Eliades, , & A. Konstantinidis, Android malware detection with unbiased confidence guarantees, *Neurocomputing*, vol. 280,no. 2018, pp.3-12, Mar ,2018.
- [22] O. Somarriba, and U. Zurutuza, A collaborative framework for android malware detection using DNS & dynamic analysis. In *2017 IEEE 37th Central America and Panama Convention (CONCAPAN XXXVII)(IEEE)* , Managua, Nicaragua , 2017,pp. 1-6.
- [23] M.R. Khan, S. Dubey, and R.C. Tripathi, Network Traffic Based Detection of Repackaged Android Apps via Mobile Fog Computing, *International Journal of Future Generation Communication and Networking*, vol 14, no. 1, pp.2824-2838, May , 2021.
- [24] F. Tong, and Z. Yan, A hybrid approach of mobile malware detection in Android, *Journal of Parallel and Distributed computing*, vol. 103, no. 2017, pp.22-31, May ,2017.
- [25] M.K. Alzaylaee, S.Y. Yerima, and S. Sezer, Emulator vs real phone: Android malware detection using machine learning. In *Proceedings of the 3rd ACM on International Workshop on Security and Privacy Analytics*, Scottsdale, Arizona USA ,2017, pp. 65-72.
- [26] M. Dietz, , S. Shekhar, Y.Pisetsky, A. Shu, and D.S. Wallach, Quire: Lightweight provenance for smart phone operating systems. In *USENIX security symposium*, Vol. 31, no. 2011, p. 3, August , 2011.
- [27] S. Bugiel, L.Davi, A. Dmitrienko, T. Fischer, and A.R. Sadeghi, Xmandroid: A new android evolution to mitigate privilege escalation attacks. Technische Universität Darmstadt, *Technical Report* , TR-2011-04, Apr, 2011.
- [28] A.T. Kabakus, and I.A. Dogru, an in-depth analysis of Android malware using hybrid techniques, *Digital Investigation*, vol. 24, no. 2018, pp.25-33, Mar,2018.
- [29] M.R. Khan, and M.K. Jain, A novel technique for detecting repacked android applications using constant key point selection-based hashing and limited binary pattern texture feature extraction, *Journal of Ambient Intelligence and Humanized Computing*,vol 12, no 2021, pp.1-12, Mar, 2021.
- [30] Z. Yuan, Y. Lu, and Y. Xue, Droiddetector: android malware characterization and detection using deep learning, *Tsinghua Science and Technology*, vol. 21 no. 1, pp.114-123, Feb, 2016.
- [31] W. Wang, M. Zhao, and J. Wang, Effective android malware detection with a hybrid model based on deep autoencoder and convolutional neural network, *Journal of Ambient Intelligence and Humanized Computing*, vol. 10, no. 8, pp.3035-3043, Aug, 2019.
- [32] H.J. Zhu, T.H., Jiang, B. Ma, Z.H. You, W.L. Shi, and L. Cheng, HEMD: a highly efficient random forest-based malware detection framework for Android, *Neural Computing and Applications*, vol 30 no. 11, pp.3353-3361, Dec, 2018.
- [33] M. Fan, J. Liu, X. Luo, K. Chen, Z. Tian, Q. Zheng, and T. Liu, Android malware familial classification and representative sample selection via frequent subgraph analysis, *IEEE Transactions on Information Forensics and Security*, vol 13, no. 8, pp.1890-1905, Feb, 2018.
- [34] D. Arp, M.Spreitzenbarth,M. Hubner, H. Gascon, K. Rieck, and C.E.R.T. Siemens, Drebin: Effective and explainable detection of android malware in your pocket. In *Ndss*, Vol. 14,no. 2014, pp. 23-26, Feb, 2014.
- [35] M. Zhang, Y.Duan, H. Yin, and Z. Zhao, Semantics-aware android malware classification using weighted contextual api dependency graphs. In *Proceedings of the 2014 ACM SIGSAC conference on computer and communications security*, Scottsdale, Arizona USA,2014, pp. 1105-1116.
- [36] T. Hsien-De Huang, and H.Y. Kao, R2-d2: Color-inspired convolutional neural network (cnn)-based android malware detections, In *2018 IEEE International Conference on Big Data (Big Data)*, New York, USA, 2018, pp. 2633-2642.
- [37] M. Yang, and Q. Wen, detecting android malware by applying classification techniques on images patterns, In *2017 IEEE 2nd International Conference on Cloud Computing and Big Data Analysis (ICC-CBDA)*, Chengdu, China ,2017, pp. 344-347.
- [38] A. Karimi, and M.H. Moattar, Android ransomware detection using reduced opcode sequence and image similarity, In *2017 7th International Conference on Computer and Knowledge Engineering (ICCKE)* , Mashhad, Iran., 2017, pp. 229-234.
- [39] E. Salahat, and M. Qasaimeh, Recent advances in features extraction and description algorithms: A comprehensive survey. In 2017 IEEE international conference on industrial technology (ICIT) , Toronto, ON, Canada,2017 , pp. 1059-1063.
- [40] H. Bay, T. Tuytelaars, and L. Van Gool, Surf: Speeded up robust features. In *European conference on computer vision*, Springer, Berlin, Heidelberg, 2006, pp. 404-417.
- [41] E. Rosten, and T. Drummond, Machine learning for high-speed corner detection. In *European conference on computer vision* , Springer, Berlin, Heidelberg, 2006 , pp. 430-443.
- [42] K. Lim, N.Y Kim, Y. Jeong, S.J. Cho, S. Han, and M. Park, Protecting Android Applications with Multiple DEX Files Against Static Reverse Engineering Attacks, *Intelligent Automation and Soft Computing*, vol. 25 , no. 1, pp.143-154, Mar, 2019.
- [43] N. Ali, K.B. Bajwa, R. Sablatnig, S.A. Chatzichristofis, , Z. Iqbal, , M. Rashid, and H.A. Habib, A novel image retrieval based on visual words integration of SIFT and SURF, *PloS one*, vol. 11, no. 6, p. e0157428, Jun, 2016.

A Hybrid Heuristic for a Two-Agent Multi-Skill Resource-Constrained Scheduling Problem

Meya Haroune¹

Université de Tours
Laboratoire d'Informatique
Fondamentale et Appliquée
de Tours (LIFAT) ROOT ERL-CNRS 7002
Université de Nouakchott Al-Aasriya

Cheikh Dhib²

Institut Supérieur du Numérique
Nouakchott, Mauritanie

Emmanuel Néron³

Université de Tours
Laboratoire d'Informatique
Fondamentale et Appliquée
de Tours (LIFAT) ROOT ERL-CNRS 7002
France, Tours

Ameur Soukhal⁴

Université de Tours
Laboratoire d'Informatique
Fondamentale et Appliquée
de Tours (LIFAT) ROOT ERL-CNRS 7002
France, Tours

Hafed Mohamed Babou⁵

École Supérieure de Polytechnique
Unité de Recherche Intelligence Artificielle
Nouakchott, Mauritanie

Farouk Mohamedade Nanne⁶

Université de Nouakchott Al-Aasriya
Unité de Recherche Calcul Scientifique,
Informatique et Science de Données
Nouakchott, Mauritanie

Abstract—This paper addresses an industrial case of the two-agent scheduling problem with a global objective function. Each agent manages one or several projects and competes with another agent for the use of common multi-skilled employees. There is a pool of employees, each of which can perform a set of skills with heterogeneous performance levels. The objectives of the two agents are both to minimize the total weighted tardiness of its tasks. Furthermore, we assume that some constraints (soft constraints) can be violated when there is no feasible schedule for the problem. Thus, the global objective function minimizes the constraint violations by reducing the undesirable deviations in the soft constraints from their respective goals. The overall objective is to find a schedule that minimizes both agents objective functions (local objectives) and the global objective function. We provide a mixed-integer goal programming (MIGP) formulation for the problem. In addition, we present a hybrid algorithm combining an exact procedure, a greedy heuristic, and a genetic algorithm to find an approximate Pareto solution set. We compare the performance of the hybrid algorithm against the corresponding MIGP formulation with simulated instances derived from real-world instances.

Keywords—Two agents; multi-skilled employees; multi-project scheduling; hybrid genetic algorithm; MIGP

I. INTRODUCTION

In the last decade, multi-project scheduling studies have introduced a new environment where different agents (local decision makers) are involved in the scheduling process. This is very common in some real-life situations, where tasks to be processed belong to different subsets and are subject to different performance measures. Such a situation can be encountered in an organization that deals with multiple projects for which the customers do not have the same requirements. For instance, some customers may be more demanding on delay, some on cost, and others on both at the same time, etc. To deal with these different requirements, new extensions of multi-project scheduling problems have been introduced, in which at least one performance measure is applied on some

tasks and not on the whole set. In addition, subsets of tasks compete for the use of common processing resources, which can create conflicts. These kinds of problems are called multi-agent scheduling problems [1].

The authors in [2] and [1] were pioneers who introduce the multi-agent concept into scheduling problems. Particularly in the two-agent scheduling model, two agents want to perform their respective tasks on common processing resources. Each agent has its own subset of tasks, which is entirely distinct from the subset of the other agent, and wants to optimize some scheduling criterion that depends on its tasks only. The goal is to determine the best compromise solutions that satisfy the agents' criteria.

This paper studies, to our knowledge for the first time, a model integrating the concept of the two-agent scheduling and multi-skill project scheduling. The two agents compete on the usage of common multi-skilled employees. Each agent manages one or more software projects that must be carried out simultaneously and completed within a fixed horizon (consecutive weeks). Each project consists of a set of independent, preemptive tasks; and each task belongs to one of the agents.

Each task is associated with a release date, a due date, and a penalty value must be paid for each week of delay. These release and due dates are negotiated with the final client and are contractually fixed. Thus, the non-respect of one of these due dates may lead to the payment of penalties. Our aim is to reduce these penalties. We consider that each task needs exactly one skill and must be performed by one employee who possesses the corresponding skill with an efficiency level. Furthermore, each task has a nominal load which corresponds to a theoretical time needed to perform this task. The nominal load for each task may be compressed according to the efficiency level of the employee in charge of that task.

There is a pool of multi-skilled employees with known weekly availability. Each employee may be involved in more

than one project at the same time with a maximum quota (percentage of time) allotted to each project. These quotas are considered here as variables and need to be calculated by the procedure of scheduling.

Furthermore, we consider that some constraints (soft constraints) can be violated when there is no feasible schedule for the problem. The global objective function seeks to minimize these constraint violations by reducing the undesirable deviations in the soft constraints from their respective goals. The objectives of the two agents are both to minimize the total weighted tardiness of its tasks.

The problem under study is a general instance of the uniform parallel machines scheduling problem with preemptions and release dates $Qm | r_j, pmtn | \sum w_j T_j$, which is known to be \mathcal{NP} -hard ([3]). However, we focus here on the case with more than one agent with different objective functions and resources with heterogeneous skills, which obviously increases the difficulty of solving an instance considerably. Our goal in this paper is to design effective heuristics that are able to generate a good approximation of the Pareto set.

The rest of the paper is organized as follows. The next section reviews the relevant literature on the two-agent scheduling problem. Section III describes the addressed problem in more detail. Sections IV and V present a mixed-integer goal programming (MIGP) formulation and heuristic approaches, respectively. Afterward, section VI-B5 presents the results of experiments conducted to analyze the performance of the proposed methods, and then Section VII concludes and presents future works.

II. LITERATURE REVIEW

This section presents the literature related to two scheduling topics that have been addressed so far separately: the multi-skill resource-constrained project scheduling problem (MS-RCPSP) and the multi-agent scheduling problem. In Section II-A, we discuss multi-skill project scheduling studies that focused on the multi-project environment. In Section II-B, we discuss multi-agent scheduling studies that specifically interested in the case of two agents competing on parallel machines. A synthesis of the reviewed literature is given in Section II-C.

A. Multi-Project Multi-Skill Resources-Constrained Project Scheduling Problems

The multi-skill Resources-Constrained Project Scheduling problem (MS-RCPSP) is an extension of the well-known Resource-Constrained Project Scheduling Problem (RCPSP), whereby multi-skilled resources (human resources or multipurpose machines) are involved. This multi-skill RCPSP extension focuses more on the particularities of human resources, such as the skills they master and sometimes the level of effectiveness in exercising those skills.

The author in [4] were the pioneers who introduced multi-skill resources into the project-scheduling field. Since then, many researchers have focused their studies on this problem considering many properties and optimizing various objectives. Particularly, most of the studies focused on MS-RCPSP merely assume that all tasks to be scheduled belong to the same

project. In this review, we focus on a multi-project setting, the reader interested in the mono-project case is referred to the papers by [5], [6].

The MS-RCPSP has been considered in a multi-project environment. The author in [7] consider a multi-project setting with heterogeneous skill resources and learning effect. The concept of learning effect means that the efficiency of resources will increase by doing more. The objective function in their study minimizes outsourcing costs. The authors in [8] and [9] considered resources with heterogeneous skills that influence the speed of work of resources. The author in [8] subdivided projects into work packages, with earliest and latest start periods associated with projects. However, [9] considered earliest and latest start periods for tasks. Thus, each task has exactly one predecessor task linked with it by maximum and minimum start-to-start time lags. The objective functions in both studies minimize the costs associated with internal and external resource usage. The author in [10] extended the same model by considering a stochastic setting. The author in [11] considered heterogeneous skills and assumed that the efficiency levels of skills may increase or decrease task duration. The objective is to assign to each project a subset of resources (team), with each team member can be assigned to several projects at a time. The author in [12] focused on a multi-objective version for project selection and scheduling problem, that includes heterogeneous skills, variable capacities over time, learning and forgetting effects. The joint problem of project selection and scheduling consists in selecting then scheduling an optimal portfolio of projects among several available projects. The objectives of the authors are the maximization of the economic gains of the selected projects and the maximization of the efficiency increase of the resources due to learning effects. The author in [13] developed a similar model for multi-project scheduling and multi-skilled staff assignment for IT product development. The objective functions they considered consist of maximizing the efficiency gain and minimizing the product development cycle time and costs. The author in [14] investigated a roughly similar model with uncertainty and learning effect. They assumed that each task requires several skills with a minimum level per skill and its processing time is related to resource efficiency. The book of [15] covers three versions for multi-project scheduling, namely project selection and scheduling, workforce assignment, and resource leveling. In the recent paper of [16], an integrated model of multi-mode and multi-skill project scheduling problem is considered. The multi-skilled resources have different skill levels, resulting in different processing times for the same task. The author focused on the minimization of the total makespan of projects.

B. Two-Agent Scheduling Problems

To better position our work more clearly in the multi-agent scheduling literature, we decide to limit our review to the related literature, which topics may be classified into (1) two-agent single-machine scheduling problems and (2) two-agent scheduling problems in a parallel machine environment.

1) *Two-Agent Single-Machine Scheduling Problems*: There is an enormous amount of literature investigating two-agent single-machine scheduling problems. Because of the large amount of literature, we only discuss in this section those

including due date-based objective functions. Agnetis et al ([1]) studied several scenarios for different combinations of objective functions of the two agents involving a single machine. The problems addressed consist in minimizing the value of one agent, while maintaining the objective value of the other agent below or at a fixed level. The objective functions they considered include the maximum of regular functions (f_{max}), number of late jobs ($\sum U_j$), and total weighted completion times ($\sum w_j C_j$). The author in [17] deal with a similar model with the goal to minimize the total completion time ($\sum C_j$) of one agent with the restriction that the number of tardy jobs ($\sum U_j$) of the other agent cannot exceed a given number. The author in [18] addressed several two-agent single-machine problems consisting in minimizing the total weighted completion time of one agent, subject to an upper bound on the value of the other agent, which may be: total weighted completion time, maximum lateness, and maximum completion time. More recently, [19] considered a similar model with an objective to minimize the weighted number of tardy tasks ($\sum w_j U_j$) of the first agent, subject to an upper bound on the weighted number of tardy jobs of the second agent. The author in [20] proposed a model similar to the models above with the objective to minimize the total weighted late tasks of one agent, while keeping the value of the total completion time of the other agent lower than or equal to a given value. The authors of [21] addressed a two-agent single-machine scheduling problem with learning effects where the objective is to minimize the total tardiness ($\sum T_j$) of the first agent, subject to an upper bound on the maximum tardiness (T_{max}) of the second agent. The author in [22] extended the model of [17] to the case with learning effect. The objective was the minimization of the total weighted completion time of the first agent with the restriction that no tardy job is allowed for the second agent. The author in [23] considered a two-agent single-machine scheduling problem with assignable due dates. The goal is to assign a due date from a given set of due dates and a position in the sequence to each task so that the weighted sum of the objectives of both agents is minimized. The authors considered several combinations of the objectives, which include the maximum lateness, total (weighted) tardiness, and total (weighted) number of tardy tasks. In [24], the author extended the same model by minimizing the objective of the first agent with an upper bound on the value of the objective of the second agent. The author in [25] considered unit processing time tasks and a common due date (see Section II-B2). The author in [26] tackled a two-agent single-machine scheduling model that considers setup times between agent tasks. The authors considered several combinations of the objectives: the maximum lateness, the total (weighted) completion time, and the (weighted) number of tardy tasks. The author in [27] considered the same setting with the objective to minimize the total weighted completion time of the first agent subject to an upper bound on the makespan of the second agent. The author in [28] assumed that tardy a task incurs a tardiness penalty cost which can be avoided by compressing the processing time of some tasks, which includes an additional cost. The objective of each agent is to minimize the total tardiness penalty cost plus the total compression cost. The authors considered two single-machine scheduling problems. The first problem is to minimize the weighted sum of the objectives of the two agents. The second problem is to minimize the objective of one agent with a constraint on the value of the objective of the other

agent.

2) *Two-Agent Scheduling on Several Machines*: Considering the above literature, it can be seen that studies including a parallel-machine environment are relatively limited. More precisely, most of the studies on the two-agent parallel-machine scheduling problem focused on identical parallel machine environment. The author in [29] were the first to consider this setting, where the objective of one agent is to minimize the makespan and that of the other is to minimize the total completion time. The author in [30] studied two models of two-agent scheduling on identical machines where the goal is to minimize the makespan and the total completion time of one agent respectively, subject to an upper bound on the makespan of the other agent. The author in [31] interested in a similar model with the goal to minimize the total weighted completion time of the first agent, subject to an upper bound on the value of the makespan of the second agent. The author in [32] tackled also a similar model with the goal to minimize the makespan of the first agent, subject to an upper bound on the makespan of the second agent. The author in [33] studied several two-agent scheduling problems for identical parallel machines with preemption and release dates for either one set or both sets of tasks. The objective functions they considered are the total (weighted) completion time, the number of tardy tasks, the total tardiness, the maximum lateness, and a regular function of type f_{max} . The author in [34] considered a two-agent setting with a single machine or two identical machines in parallel. The processing times of the tasks of one agent are compressible at an additional cost. The authors considered several different objective functions: the regular function f_{max} , the total completion time plus compression cost, the maximum tardiness plus compression cost, the maximum lateness plus compression cost, and the total compression cost subject to deadline constraints. The author in [35] studied a two-agent parallel-machine scheduling model with the assumption that a task can be rejected, which incurs a penalty. The objective is to minimize the sum of the scheduling cost of the accepted tasks and the total rejection penalty of the rejected tasks. The authors considered several combinations of objectives: the makespan, the total completion time, the maximum lateness, and the weighted number of tardy tasks. The author in [36] studied a two agent scheduling problem with deteriorating effect on bounded parallel batching machines. The objective is to minimize the makespan of one agent with the constraint that the makespan of the other agent is no more than a given threshold. The author in [37] study a scheduling problem for concurrent jobs on identical parallel machines. It deals with an interfering multi-agent scheduling problem. New complexity results have been developed when the jobs are of identical durations. Some problems are shown to be polynomial where exact solution algorithms are developed and others are shown to be \mathcal{NP} -hard.

To the best of our knowledge, very few studies focused on a setting of two agents competing on uniform parallel machines. The author in [38] considered identical processing time tasks where the goal is to minimize at the same time a general cost function associated with the first agent and the makespan of the other agent. In [39], the author addressed the same model with the goal to minimize two maximum functions associated with the two agents. The author in [25] developed a single machine, and parallel (both identical and uniform) machine

settings. They discussed the case where the tasks have identical processing times and a common due date. They focused on minimizing the total weighted earliness–tardiness of the first agent, subject to an upper bound on the maximum weighted deviation from the common due date of the tasks of the second agent.

There exist at least two studies that tackled the case of two competing agents on unrelated parallel machines. The author in [40] considered a Just-in-Time setting where the objective of the first agent is to maximize the weighted number of its just-in-time tasks, while the objective of the second agent is either to maximize its maximum gain from its just-in-time jobs or to maximize the weighted number of its just-in-time jobs. The author in [41] focused on the objective of minimizing the total completion time of the tasks of one agent, while keeping the weighted number of tardy tasks of the other agent within a given limit.

C. Synthesis

In this paper, we study an integrated multi-agent scheduling and multi-skill project scheduling problem with many particularities. To the best of our knowledge, this problem has never been studied in the literature. The novelty of our model is related also to several particularities stem from the preferences of the managers. For instance, we consider minimum and maximum loads associated with each task. Also, we consider that each employee, should not exceed a fixed number of different tasks over a given week. We consider a preemptive MS-RCPSp problem with a multi-project setting and heterogeneous skill levels. In light of the existing literature on multi-project multi-skill RCPSp, it is noticed that the studies including those two features are very limited.

Table I presents a synthesis of the studies reviewed above. The first column indicates the paper. The second column provides the characteristics of mutli-skilling: “#SK” indicates the number of skills required by the task, “HS” for heterogeneous skills and “ML” indicates if a minimum level of skill is required to perform the task. The third column indicates some multi-agent features presented according to several sub-columns. Sub-column “M” indicates the number of machines. Sub-column ‘E’ describes the parallel machine environment (“Pm” for identical machines (Qm” for uniform machines “Rm” for unrelated machines). Sub-column “O” indicates that the objective value of one agent is constrained under an upper bound. Sub-column “C” means that the objective function is a combination of the agent’s objectives functions. Sub-column “P” means that the goal is to enumerate the entire Pareto frontier. The fourth column “Objective” shows the objective function. Followed by the last column, which indicates if the paper considers other specific characteristics.

For each paper in Table I, we use the following abbreviations to indicate the objective considered. OC: outsourcing costs; EC: External cost; ATS: average team size; SEG: skill efficiency gain; PDCT: product development cycle time and costs; $w_j E$: weighted numbers of just-in-time tasks, G : gain from just-in-time tasks; $w_j C_j$: weighted total completion time; C_{max} : project duration (makespan); f_{max} : maximum of regular functions; L_{max} : maximum lateness; $w_j U_j$: weighted number of tardy tasks.

III. PROBLEM DEFINITION AND NOTATIONS

There is a set $K = \{k_1, \dots, k_L\}$ of L projects that must be completed before a common due date (horizon). There are two competing project managers (agents), called A and B , each has a disjoint subset of projects. Each project k_l is broken down into a set of independent, preemptive tasks; and each task belongs to one agent. As explained earlier, we do not consider precedence constraints either between projects or between tasks. As the tasks are independent, we suppose that all the tasks are numbered from 0 to $J + 1$, where the 0th and the $n + 1$ th tasks are dummy indicating the start and end of projects, respectively. The subset of tasks of agents A and B are denoted by $N_A = \{n_1, \dots, n_{J_A}\}$ and $N_B = \{n_{J_A+1}, \dots, n_J\}$, respectively. The time unit is the half-day.

For each task n_j , there is a nominal load c_j (expressed in man-days), a release date r_j (given in weeks), and a minimum load denoted c_j^{\min} (expressed in half-days) to quantify the minimum degree of realization of this task per week. The minimum load of tasks per week allows modeling some tasks that cannot be interrupted during more than one week. In the case where the employee assigned to a task n_j works on that task during a given week, he or she should perform at least its minimum load c_j^{\min} . Furthermore, In each week, the employee assigned to task n_j must not exceed its maximum load c_j^{\max} . We want to avoid loss of time due to changing context of employees, which is required when changing from one task to another. Thus, during each week, the number of different tasks on which an employee is working is less than a given value b (fixed for all the projects and all the employees).

Let $E = \{e_1, \dots, e_I\}$ be a set of I multi-skilled employees working in the company. Every employee has an availability per week (a working time known in advance) ranging from 0 to 10 half-days. We refer by $D_{i,s}$ the availability of employee e_i during week h_s , where $h_s \in H$. Employees are allocated to different projects with maximum percentages of time (quotas). As mentioned earlier, these quotas must be determined by the scheduling procedure. Once determined, they must be respected during each week of the horizon. In other words, during each week h_s , any employee e_i assigned to project k_l cannot spend on this project more than $D_{i,s} \times Q_{i,l}$, where $Q_{i,l}$ is the quota of employee e_i on project k_l . Each employee can work on only one task at a given time frame.

In our model, once a task is assigned to an employee with the required skill, it remains so until its accomplishment. The capabilities of performing tasks by resources are represented by a binary skill matrix denoted by m , where $m_{j,i} = 1$ if employee e_i masters task n_j , and $m_{j,i} = 0$ otherwise. It means not every employee can be assigned to a task. Furthermore, we assume that several employees may have different levels of efficiency for the same skill. Since each task requires only one skill, the skill level of the employee is directly associated with the task. The manager estimates the employees’ efficiency level according to the standard classification of expertise level: junior, middle and senior. Based on these estimations, we assign an efficiency coefficient equal to 0, 0.5, and 1 to a junior, middle and senior, respectively. This coefficient is a ratio of an employee’s actual processing time to perform the task against the theoretical processing time (nominal load) needed to complete the corresponding task. Thus, to consider

the employee's efficiency level in the processing time of task calculation, a simple linear formula is assumed between the task nominal load and the employee assigned to it. We apply this formula to convert the task nominal load (c_j) to duration (processing time, $p_{j,i}$) according to the efficiency level ($v_{j,i}$) of the employee: $p_{j,i} = (2 - v_{j,i})c_j$. Since the nominal load of the task is given in number of days, we multiply it by 2 to convert it into half-days. For example, a task that requires a java developer and 2 days to be performed, it can be done by a senior developer in 2 half-days (i.e. half of the time), and by a junior developer in 4 half-days (the actual time).

The constraints on the minimum load of tasks per week and on the number of different tasks on which an employee is working are soft constraints imposed by the manager to increase the productivity of the employees.

For an effective schedule, these soft constraints should be taken into account. However, the manager allows the soft constraints to be violated when there is no feasible schedule for the problem. The solution approach must minimize these constraint violations by reducing the undesirable deviations in the soft constraints from their respective goals. We introduce a global objective function to penalize these constraint violations. All the other constraints (also called hard constraints) must be respected by the proposed solution.

The objective functions considered here are as follows. Let f^A and f^B be the objective functions of agents A and B , respectively. Each of the agents wants to minimize the total weighted tardiness of its tasks denoted by $f^X = \sum_{j \in N_X} w_j T_j$, where $X \in \{A, B\}$, T_j is the number of weeks of task n_j tardiness and w_j is the penalty cost for this task. Note that if task n_j takes at least one half-day of week ($d_j + 1$) before its completion time, it is late by one week ($T_j = 1$). Furthermore, the soft constraints are addressed as goals to be reached, and the global objective is to get as close as possible to these goals. We consider a global objective function that

consists to minimize the violations of these constraints. This objective function is defined by $f^G = \sum_{j=1}^J \sum_{s=r_j}^H \alpha u_{j,s}^- + \sum_{i=1}^I \sum_{s=1}^H \beta o_{i,s}^+$, where $u_{j,s}^-$ is the deviation below c_j^{\min} , $o_{i,s}^+$ is the deviation above b , and α and β are problem parameters stem from the preferences of project managers on soft constraints. Between the two soft constraints, the minimum load constraint is slightly more important than the other soft constraint. Hence, α is slightly higher than β . The problem is to find a schedule that minimizes at the same time the objective functions of both agents as well as the global objective function.

Following the conventional three-field notation introduced by [42] and extended by [43], this problem may be denoted by: $Qm \mid CO - GA, r_j, pmtn \mid f^G, f^A, f^B$, where $CO - GA$ denotes a problem of disjoint subsets competing with a global objective.

A solution consists of two parts: the first is to specify a suitable allocation of employees to projects, and the second is to determine a schedule of tasks for each employee to complete within the planning horizon H . Note that a schedule is defining by a suitable assignment of employees to task and the load (i.e. the number of half-days) that each employee has to perform of each of its tasks during each week. We are interested in determining a good approximation of the Pareto frontier.

TABLE 1. SYNTHESIS OF THE REVIEWED LITERATURE

Auteurs	Multi-skill features			Multi-agent features				Objective	Other
	# Sk	HS	ML	# M	E	O	C		
This paper	1	•	•	> 1	Q_m				Multi-project, preemption, minimum and maximum loads, soft constraints
[16]	1			1		•			Multi-project, release dates for projects, multi-mode
[19]	≥ 1	•			Q_m				Multi-project, multi-objective, uncertainty, learning effect
[28]	≥ 1	•		> 1	P_m	•			Controllable processing times
[36]	≥ 1					•			Deteriorating effect
[10]	≥ 1			1		•			Multi-project, stochastic concept, internal and external resources
[27]	1			1		•			Setup time
[13]	1			> 1	P_m	•			Multi-objective, Learning and forgetting effects
[35]	1			> 1	P_m	•			Task rejection
[31]	1			> 1	P_m	•			Setup times
[26]	1	•		1		•			Multi-project
[11]	1			> 1	P_m	•			Learning effects
[32]	1			> 1	P_m	•			Controllable processing times
[21]	1			< 2	P_m	•			Assignable due dates
[34]	1			1	P_m	•			Unit processing time tasks, common due date
[24]	1			1 and > 1	$Q_m, P_m, P1$	•			Internal and external resources, overtime
[25]	≥ 1			> 1	P_m	•			Assignable due dates
[30]	≥ 1	•				•			Learning effect
[9]	≥ 1	•		1		•			Multi-project, Dynamic competencies, time-dependent capacities
[23]	1			1		•			Internal and external resources, overtime
[22]	1			1		•			machines, Preemption
[12]	≥ 1	•	•	> 1	P_m	•			Multi-project, Learning effects, external resources
[8]	≥ 1	•		1	P_m	•			Identical processing time
[33]	1			> 1	P_m	•			Identical processing time
[18]	1			1		•			Just-in-time
[29]	≥ 1	•		> 1		•			
[7]	≥ 1			1		•			
[17]	1			1		•			
[1]	1			1		•			
[38]	> 1			> 1	Q_m	•			
[39]	> 1			> 1	Q_m	•			
[40]	> 1			> 1	R_m	•			

Table II presents the notation of the problem parameters used throughout this paper.

TABLE II. NOTATIONS AND PARAMETERS OF THE PROBLEM

General data	
H	project planning horizon
X	index of agent X , $X \in A, B$
K	set of projects, $K = \{k_1, \dots, k_L\}$, $ K =L$
E	set of employees, $E = \{e_1, \dots, e_i\}$, $ E =I$
N	set of tasks, $N = \{n_1, \dots, n_j\}$, $ N =J$
N_X	set of agent X 's tasks, $ N_X = J_X$ and $J = J_A + J_B$
N_l^k	set of tasks of project k_l
WL_l^k	nominal load of project k_l in skill
$Z_l^k = \{z_k \ k \in \{1, \dots, K\}\}$	required skills to complete k_l project
Tasks data	
c_j	nominal load of task n_j (measured in man-day)
r_j	release date of task n_j (given in number of weeks)
d_j	due date of task n_j (given in number of weeks)
c_j^{\max}	maximum load of task n_j (given in half-day)
c_j^{\min}	minimum load of task n_j (given in half-day)
w_j	penalty cost of task n_j per week
F_j	completion date of task n_j (given in number of weeks)
T_j	tardiness of task n_j (given in number of weeks)
E_k	set of employees mastering skill z_k
Employees data	
$D_{i,s}$	availability of employee e_i during week h_s (given in half-days)
b	number of different tasks on which every employee can work during each week
$M_{i,k}$	1 if employee e_i masters skill z_k required by project k_l , and 0 otherwise
Data on tasks and employees	
$v_{j,i}$	employee e_i 's efficiency level for task n_j
$p_{j,i}$	processing time of task n_j according to the efficiency level of employee e_i (in half-days)
$m_{j,i}$	equal to 1 if employee e_i masters task n_j , and 0 otherwise
\bar{v}_j	average efficiency of employees mastering task n_j

IV. MIXED-INTEGER GOAL PROGRAMMING FORMULATION

A. Variables

- 1) $x_{j,i}$ –a binary variable equals 1 if employee e_i is assigned to task n_j and equals 0 otherwise.
- 2) $y_{j,i,s}$ –an integer variable (ranging from 0 to 10) equal to the number of half-days performed of task n_j by employee e_i during week h_s .
- 3) $z_{j,i,s}$ –a binary variable equal to 1 if $y_{j,i,s}$ is greater than 0, and equal to 0 otherwise.
- 4) F_j –the completion time of task n_j .
- 5) T_j –the tardiness of task n_j .
- 6) $u_{j,s}^+$ –deviation variable above c_j^{\min} associated to task n_j and week h_s .
- 7) $u_{j,s}^-$ –deviation variable below c_j^{\min} associated to task n_j and week h_s .
- 8) $o_{i,s}^+$ –deviation variable above b associated to employee e_i and week h_s .
- 9) $o_{i,s}^-$ –deviation variable below b associated to employee e_i and week h_s .
- 10) $Q_{i,l}$ –maximum quota of employee e_i on project k_l (percentage of time).

[]

B. Constraints

$$\sum_{i=1}^I x_{j,i} = 1, \quad j = 1, \dots, J \quad (1)$$

$$x_{j,i} \leq m_{j,i}, \quad j = 1, \dots, J; \quad i = 1, \dots, I \quad (2)$$

$$\sum_{s=r_j}^H y_{j,i,s} = p_{j,i} \cdot x_{j,i} \quad j = 1, \dots, J; \quad i = 1, \dots, I \quad (3)$$

$$\sum_{i=1}^I y_{j,i,s} \leq \sum_{i=1}^I c_j^{\max} \cdot x_{j,i}, \quad j = 1, \dots, J; \quad s = r_j, \dots, H \quad (4)$$

$$\sum_{i=1}^I z_{j,i,s} \leq \sum_{i=1}^I x_{j,i} \quad j = 1, \dots, J; \quad s = r_j, \dots, H \quad (5)$$

$$\sum_{i=1}^I 10 \cdot z_{j,i,s} \geq \sum_{i=1}^I y_{j,i,s}, \quad j = 1, \dots, J; \quad s = r_j, \dots, H \quad (6)$$

$$\sum_{j=1}^J y_{j,i,s} \leq D_{i,s}, \quad i = 1, \dots, I; \quad s = 1, \dots, H \quad (7)$$

$$\sum_{j \in N_l} y_{j,i,s} \leq Q_{i,l} \cdot D_{i,s}, \quad i = 1, \dots, I; \quad s = 1, \dots, H; \quad l = 1, \dots, L \quad (8)$$

$$\sum_{i=1}^I y_{j,i,s} \geq \sum_{i=1}^I c_j^{\min} * z_{j,i,s}, \quad j = 1, \dots, J; \quad s = r_j, \dots, H \quad (9)$$

$$\sum_{j=1}^J z_{j,i,s} \leq b, \quad i = 1, \dots, I; \quad s = 1, \dots, H \quad (10)$$

$$F_j \geq \sum_{i=1}^I s \cdot z_{j,i,s}, \quad j = 1, \dots, J; \quad s = r_j, \dots, H; \quad (11)$$

$$F_j - T_j \leq d_j, \quad i \quad F_j, T_j \geq 0 \quad (12)$$

C. Soft Constraints

According to the problem description, constraints (9) and (10) are soft constraints that can be violated when it is not possible to obtain a feasible schedule. Therefore, we incorporate the possibility of relaxing these soft constraints by adding deviation variables in the formulation. Meanwhile, these deviation variables are calculated and added to the objective function as penalties. After adding the deviation variables, the soft constraints in equations (9) and (10) became, respectively:

$$\sum_{i=1}^I y_{j,i,s} - u_{j,s}^+ + u_{j,s}^- = \sum_{i=1}^I c_j^{\min} * z_{j,i,s}, \quad j = 1, \dots, J; \quad s = r_j, \dots, H \quad (13)$$

$$\sum_{j=1}^J z_{j,i,s} - o_{i,s}^+ + o_{i,s}^- = b, \quad i = 1, \dots, I; \quad s = 1, \dots, H \quad (14)$$

D. Objective Functions

The three different objective functions are listed as follows.

Global Objective: the goal associated to constraint (13) is to avoid as much as possible that, on a given week, the employee assigned to a task perform of this task less than its minimum load. Because of that, only the negative deviation from this goal is minimized in the following equation.

$$\text{Min} \sum_{j=1}^J \sum_{s=r_j}^H u_{j,s}^- \quad (15)$$

The goal associated to constraint (14) is to limit the loss of employee time due to switching between tasks. To this end, no employee should work on more than the maximum number of tasks per week. Therefore, only the positive deviation from this goal is minimized in the following objective function.

$$\text{Min} \sum_{i=1}^I \sum_{s=1}^H o_{i,s}^+ \quad (16)$$

After incorporating these goals, the achieving global objective function can be written as follows.

$$\text{Min} \sum_{j=1}^J \sum_{s=r_j}^H \alpha u_{j,s}^- + \sum_{i=1}^I \sum_{s=1}^H \beta o_{i,s}^+ \quad (17)$$

Local Objectives the agents' objective functions in equations 18 and 19 seek to minimize the total weighted tardiness of their tasks.

$$\text{Minimize} \sum_{j=1}^{J_A} w_j T_j \quad (18)$$

$$\text{Minimize} \sum_{j=J_A+1}^J w_j T_j \quad (19)$$

V. HEURISTIC ALGORITHMS

Due to its complexity, an exact resolution of the problem is very difficult within a reasonable computation time. Therefore, we propose a hybrid algorithm combining an exact procedure, a greedy heuristic, and a genetic algorithm to find an approximate Pareto solution set. The main steps of this hybrid algorithm are as follows:

- Use a mixed integer-linear program (MILP) to set maximum quotas of employees' time on projects.
- Use a greedy heuristic to determine initial solutions.
- Apply a genetic algorithm of type NSGA-II to determine a good approximation of the Pareto frontier.

In the next sections, we detail each of the steps of the hybrid algorithm.

A. Generating Maximum Quotas

This procedure, denoted by PGQ_{exact} , is used to allocate to each project the set of employees with the necessary skills for its realization. Furthermore, it specifies the working time that each employee must not exceed per week on each of its projects. We use a simplified MILP model, which considers only the constraints on the weekly availability of employees and the workload of projects to be carried out. This MILP model uses time-indexed decision variables.

1) *Variables:* We define the integer variable $Y_{i,l,s,k}$ (ranging from 0 to 10) equal to the effective working time of employee e_i on project k_l during week h_s on skill z_k . The integer variable $Q_{i,l}$ (ranging from 0 to 10) equal to the maximum quota of employee e_i assigned to the project k_l . The integer variable $d_{i,s}^-$ (ranging from 0 to $D_{i,s}$) equal to the unused availability of employee e_i during week h_s .

2) *General Formulation:* The general formulation is given in the following.

$$\text{Minimize} \sum_{i=1}^I \sum_{s=1}^H d_{i,s}^- \quad (20)$$

s.c.

$$Y_{i,l,s,k} \leq 10 \cdot M_{i,k} \quad i = 1, \dots, I; \quad s = 1, \dots, H; \\ k = 1, \dots, K; \quad l = 1, \dots, L; \quad (21)$$

$$\sum_{k=1}^K Y_{i,l,s,k} \leq D_{i,s} \cdot Q_{i,l}, \quad i = 1, \dots, I; \quad s = 1, \dots, H; \\ l = 1, \dots, L \quad (22)$$

$$\sum_{l=1}^L \sum_{k=1}^K Y_{i,l,s,k} + d_{i,s}^- = D_{i,s}, \quad i = 1, \dots, I; \quad s = 1, \dots, H \quad (23)$$

$$\sum_{i=1}^I \sum_{s=1}^H Y_{i,l,s,k} = WL_l^k, \quad l = 1, \dots, L \quad (24)$$

The objective function (20) minimizes the sum of the unused working time (idle time), by avoiding that employees work, during each week, less than their availability. Constraint (21) guarantees that an employee must not work on a project that he or she does not master any of the skills required by that project. Constraint (22) ensures that, on any given week, no employee must exceed his or her quota on each project. Constraint (23) guarantees that an employee must not exceed his availability each week. Constraint (24) imposes that the nominal load of each project in each skill must be executed until completeness.

B. Greedy Heuristic

This algorithm employs simple priority rules and a simple heuristic to construct good initial solutions. This step is very important because a suitable assignment may help the hybrid algorithm to find an approximate Pareto solution set rapidly and effectively. This greedy heuristic returns all solutions obtained after a computation time-limited to AG_{max} .

Each initial solution is generated through two steps. The first step defines the order in which the tasks will be selected, the next step chooses, for each task, the employee who will be in charge of it among the employees mastering that task. We detail below the three steps of the greedy heuristic.

The procedure of the first step returns the list of tasks ordered in non-decreasing order of the number of employees mastering each task. For example, a task mastered by one employee should be assigned before another task mastered by two employees. This ensures that the most critical employees are not overloaded by other tasks that have multiple assignment options. In the case where two tasks are mastered by the same number of employees, the first task to be assigned is chosen according to the weighted earliest due date (WEDD) priority rule.

The procedure of the second step assigns tasks to employees. To perform this step, we proceed as follows. First, for each task n_j in the order of the list of tasks, we get the list of employees E_j mastering that task and allocated to project k_l (task n_j is part of project k_l). For each employee $e_i \in E_j$, we get $\tau_{j,l}^{tot}$ which corresponds to the total availability of employee e_i on project k_l . Then, the employee with the highest total availability will be assigned to task n_j . We note that before selecting this employee we subtract from his total availability the processing time ($p_{j,i}$) of task n_j , which ensure that he has sufficient availability to perform this task. Otherwise, if the employee does not have the required availability, the second employee in the list will be selected and so on. Finally, we update the availability of the selected employee. This means that the choice of the employee for the next task is influenced by the workload that has already been assigned.

The first and second steps of the GA generate a single solution. The remaining solutions are generated iteratively by performing simple mutation operations on the list of tasks, then repeating the second step. Algorithm (1) describes the procedure of the greedy heuristic.

C. Adaptation of the NSGA-II

NSGA-II (for Non-dominated Sorting Genetic Algorithm II) is a genetic algorithm well known as one of the most efficient and popular algorithms for solving multi-criteria optimization problems. NSGA-II method is originally proposed by [44] on the basis of NSGA proposed in [45]. We recall in the following the main ideas of this method. For more details, the reader can refer to [44]. First, individuals (solutions) are classified into a number of dominance ranks at each generation using a fast non-dominated sorting method with low computational complexity. Second, a parameter-independent partitioning method was defined by evaluating the crowding distance of individuals in the same dominance rank. Third, a selection operator was designed based on the values of

Algorithm 1 Overview of the Greedy Algorithm

```

1: Inputs:  $N$  the set of all tasks ordered in ascending order of the number of employees mastering each task;  $\tau^{tot}$  the
total availability of employees on projects
2: Output:  $\sigma$  the assignment of employees to tasks
3: for each Task  $n_j \in N$  do
4:    $k_l$  : the project to which  $n_j$  belongs
5:    $E_j$  : the list of employees mastering  $n_j$ 
6:    $BinomiaList = \phi$ 
7:   for each Employee  $e_i \in E_j$  do
8:     Add  $(e_i, \tau_{i,l}^{tot})$  to  $BinomiaList$ 
9:   end for
10:  Short  $BinomiaList$  in decreasing order of  $\tau^{tot}$ 
11:   $NotAffected = true, count = 0$ 
12:  while  $NotAffected$  do
13:    if  $BinomiaList[count][1] - p_{j,i} \geq 0$  then
14:      Assign employee  $BinomiaList[count][0]$  to task  $n_j$ 
15:       $NotAffected = false$ 
16:    end if
17:     $count++$ 
18:  end while
19:  update the availability of the employee
20:  Update list  $\sigma$ 
21: end for
22: return  $\sigma$ 

```

dominance rank and crowding distance of individuals. Finally, an elitism strategy was used to improve the convergence performance of the algorithm.

Our implementation of NSGA-II is based on the following elements. (i) the coding scheme to represent an individual, (ii) the genetic operators to generate and modify new individuals, (iii) the parameters of the algorithm (i.e. population size P_{max} , number of crossover points P_c and mutation points P_m , termination criterion G_{max}).

1) *Coding Scheme:* A chromosome (or an individual or a solution) contains a given number of genes and is divided into one or more segments. In our NSGA-II, an individual is represented by an assignment of employees to tasks, a gene corresponds to a task, a segment corresponds to the set of tasks assigned to an employee, and the length of the segment corresponds to the number of tasks in this segment. Thus, a separator (0) is used in the coding scheme of an individual to indicate the beginning and the end of each segment.

In order to understand the coding scheme, let's consider the case of two agents A and B , each of which is in charge of a single project. The projects consist of 8 tasks (n_1, \dots, n_8) to be scheduled over a two-week horizon. Three employees (e_1, e_2, e_3) can be assigned to different tasks. Each agent wants to minimize the sum of the weighted delays of its tasks. Table III shows for each task, the employee with the skill required to perform it. The gray columns represent the tasks of agent B , while the white columns represent the tasks of agent A .

TABLE III. SUITABLE EMPLOYEE ASSIGNMENTS TO TASKS

	n_1	n_2	n_3	n_4	n_5	n_6	n_7	n_8
c_j	3	5	2	6	4	4	5	3
$m_{j,1}$	1	0	0	1	1	0	0	1
$m_{j,2}$	1	1	1	1	0	0	1	1
$m_{j,3}$	0	0	1	0	1	1	1	1

Fig. 1 shows a solution coding for the previous example. In this solution, tasks n_4, n_5 are assigned to employee e_1 ; tasks n_1, n_2, n_7 are assigned to employee e_2 ; and tasks n_3, n_6, n_8 are assigned to employee e_3 . The 0 is used to separate two assignments.

2) *Initial Solutions:* Initial solutions are generated using the greedy heuristic described in Section V-B. Note that the

n_4	n_5	0	n_1	n_2	n_7	0	n_3	n_6	n_8
-------	-------	---	-------	-------	-------	---	-------	-------	-------

Fig. 1. Representation of a Solution.

number of generated solutions can be smaller than the initial population size P_{max} (P_{max} is a parameter of the NSGA-II method). In this case, we apply simple mutation operations on the elitist individuals to complete the initial population.

3) *Genetic Operators*: These operators comprise several key elements: selection, crossover, mutation, and population sorting.

Selection: it chooses among the solutions of the parent population, those that will reproduce and generate the new population (offspring) for the following iteration. To this end, we use the binary tournament method, which is one of the most common selection operators. This method randomly selects a pair of parents from the population and compares their fitness functions. If both parents are on the same front, then we compare their crowding distances, and the one with the largest crowding distance is kept. Otherwise, if the two parents are on different fronts, the solution with the best Pareto front (i.e. the one with the best value of the objective function) is kept. Both solutions of a couple are unique, but a solution can be part of several couples.

Crossover: once the selection process is done, the entire selected parents move on to the breeding stage. This is where all the parents recombine in some way to create a new population that will be used in the next genetic step (mutation). The process of combining two parents is what is often called crossover.

We adopt the two crossover operators PBX (position-based crossover) and OBX (order-based crossover) proposed by [46]. The choice of these crossover operators is made based on the study published in [47]. The authors compared the performances of 11 crossover operators with the goal to minimize the total weighted tardiness on a single machine. Their experimental results has shown the efficiency of OBX and PBX crossover operators, compared to other operators, to solve this type of scheduling problem. We detail below the different steps of these two crossover operators. Also, two corresponding examples are shown in Fig. 2 and 3.

• **OBX Operator**

- Select randomly several genes (tasks) from a parent ($P1$ for example).
- Place the selected tasks in the new solution, respecting the exact positions that they occupy in the other parent ($P2$).
- Delete the tasks that are already selected in the other parent ($P2$) to avoid repeating these tasks in the offspring ($O1$).
- Insert the remaining tasks into in the offspring, from left to right, in the order that they appear in the parent ($P2$).
- Place the remaining tasks in the offspring, from left to right, in the order that they appear in the parent ($P2$).

- By changing the roles of the parents, the same procedure can be applied to generate the offspring $O2$.
- **PBX operator**
 - Select randomly a set of tasks from a parent ($P1$ for example).
 - Place the selected tasks in the offspring ($O1$, respecting their exact positions in the parent ($P1$)).
 - Delete the tasks that are already selected in the second parent $P2$. The sequence of remaining tasks in $P2$ contains only those tasks that the offspring ($O1$) needs.
 - Place the remaining tasks in $O1$, from left to right, in the order they appear in the parent $P2$.
 - By changing the parent roles, the same procedure can be applied to generate the second offspring ($O2$).

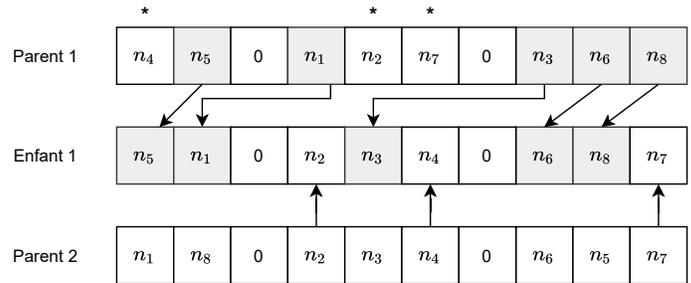


Fig. 2. OBX Crossover Example

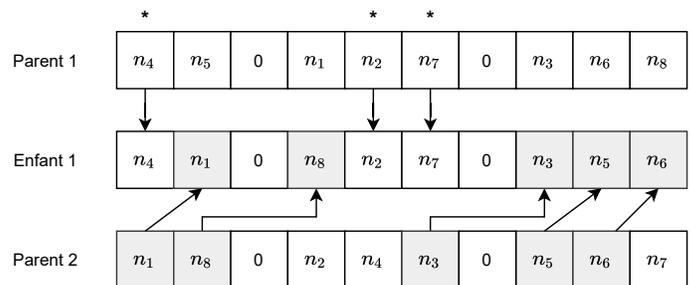


Fig. 3. PBX Crossover Example

The similarity between parent and offspring populations depends on the number of crossover points. This number (denoted by P_c) represents the number of tasks selected to put into the offspring. Preliminary tests show that a large value of P_c gives a higher probability of building solutions similar to their parents. However, a small value of P_c allows the selection of more distant solutions.

Mutation: We apply this operator at the last stage of the population generation procedure. It maintains the diversity between individuals, and therefore, avoids a premature convergence to a local optimum. The mutation operator is applied to each new solution after the crossover stage. It consists in randomly selecting two employees (e_1 and e_2 , for example) who have at least one common skill. Then, each task initially assigned to e_1 and demanding a skill mastered by e_2 will be assigned to the latter. The opposite is also applied to tasks initially assigned to e_2 . These two operations are performed P_m times, where P_m (ranging between 2 and 10.) is a parameter of the NSGA-II method.

Population Sorting: This procedure follows the same dominance sorting procedure proposed by [44]. The crowding distance is also applied.

VI. COMPUTATIONAL EXPERIMENTS

This section outlines the characteristics of the instances and the computational results of the proposed methods. Our experiments consist of three parts. In the first part, we seek to determine the size of problem instances that can be solved by the MIGP within a reasonable computation time. In the second part, we want to determine the best parameterization for the NSGA-II method. Finally, in the last part, we evaluate the performance of the heuristic methods.

Our experiments were performed on an Intel® core™ i7-1.9 GHz with 16 GB of RAM under Windows 10. All algorithms were written in Python. The CPLEX 12.8.0 solver associated with Python API was used for solving the MIGP model, using the default parameters except for the time parameter.

A. Instance Characteristics

The methods proposed in this section are experimentally validated on instances derived from real instances. AS mentioned in the introductory section, our study comes from a real scheduling case raised in an IT company. However, to perform our experiments, we could not get enough real data for confidentiality reasons. So, based on the description of the problem and the few real data given by our partner company, we generated 8 sets of instances ($T1, \dots, T8$) with 40 instances per set.

Table IV describes the characteristics of some general parameters per instance set. From left to right, the columns indicate the instance set, the project planning horizon, the number of projects, the number of employees, and the last column is the number of tasks. For each group of instances, the number of tasks per agent is chosen between $0.4xJ$; $0.5xJ$; $0.6xJ$, with J being the number of tasks. The total number of skills required to perform the tasks is ranging between 3 and 10; and each employee masters 2 to 5 skills. Each task requires one skill with a nominal load of up to 15. The release dates of 30% of tasks (randomly chosen) range between $0.5xH$ and $0.75xH$, equal to 0 for the other tasks. The due dates of 50% of tasks (also randomly chosen) are between $0.5xH$ and $0.75xH$, equal to H for the other tasks. For 30% of tasks (randomly chosen), the values of the maximum loads (resp. the minimum loads) are between 2 and the nominal loads (resp. 2 and the maximum loads). For 80% of employees, we set the availability at 10 during each week of the planning horizon. The availability of the remaining employees is between 2 and 7.

B. NSGA-II Algorithm Evaluation

We present in this section the computational results of the NSGA-II method. First, we conduct some experiments to adjust the NSGA-II parameters. Then, we perform a second test campaign to compare the performances of the two crossover operators (OBX and PBX). Finally, we conduct our last experiments to measure how much the genetic algorithm improved the results of the greedy heuristic.

TABLE IV. GENERAL PARAMETER VALUES PER INSTANCE SET

Instance set	H	L	I	J
T1	4	2	3	15
T2	4	2	3	20
T3	6	2	5	25
T4	8	2	10	50
T5	10	4	15	100
T6	14	4	20	150
T7	18	4	25	200
T8	24	4	30	250

To evaluate the quality of the returned solutions, we apply three performance metrics widely used for multi-criteria optimization problems. These metrics are the hypervolume (HV)(originally proposed by [48]), the generational distance (GD) (originally proposed by [49]), and the Pareto front size (PFS). Note that in this part of the experiments, we used 10 instances on each set of the dataset. The maximum quotas of employees on projects are initially computed by the MILP presented in Section V-A.

1) *Parameters Setting:* The NSGA-II has the following parameters to define: the population size P_{max} , number of iterations G_{max} , number of crossover points P_c , and number of mutation points P_m . The NSGA-II algorithm was run 5 times with each combination of parameters and the best values obtained are presented in Table V. To maintain test consistency, for each run, we start the genetic algorithm from the same initial solutions generated by the GH . In order to fix the number of crossover points, we performed the tests with both crossover operators. We noticed reassuringly that the best values obtained are often the same with the two operators.

TABLE V. THE VALUES OF PARAMETERS USED BY THE NSGA-II

Parameter	Range	Value
P_{max}	[100,300]	200
G_{max}	[200,3000]	1000
P_c	[1,10]	8
P_m	[1,10]	7

TABLE VI. MILP RESULTS FOR A COMPUTATION TIME LIMITED TO 5 MIN

Instance set	Number of feasible instances	Number of instances solved to optimality	Avg. time to the optimality	Avg. GAP from the optimal
T1	40	40	72.8	0%
T2	40	40	122.2	0%
T3	40	40	150.7	0
T4	40	40	194.6	0%
T5	40	38	256.8	2.5%
T6	40	35	299.2	4.5%
T7	40	32	316.5	76.7%
T8	40	29	396.9	7.6%

2) *Comparison of Crossover Operators:* Using the parameter values presented in Table V, and from the same initial solutions, we ran the NSGA-II algorithm 5 times using the OBX and PBX crossover operators. The average values of the results were calculated.

Fig. 4 shows the computed generational distances between the exact Pareto and the approximated Pareto fronts obtained with each operator. Note that, we obtain the exact Pareto fronts

TABLE VII. COMPARISON OF METHODS NSGA-II AND GH

Instance set	GPNE		NSGA-II				AG		
	PFS*	CPU*(s)	PFS	HV	GD	CPU(s)	PFS	HV	GD
T1	3.42	746.22	2.82	0.82	3.51	177.12	1.81	0.47	8.91
T2	2.91	961.86	2.13	0.76	4.43	182.72	1.54	0.43	9.32
T3	2.52	1021.98	1.89	0.73	3.87	202.98	1.95	0.38	8.12
T4	2.23	1746.78	1.63	0.68	4.43	256.10	1.42	0.35	7.25
T5	1.85	3189.21	1.51	0.64	5.14	301.45	0.26	0.28	7.89
T6	0.78	6976.81	1.51	0.62	5.14	389.95	0.42	0.18	6.91
T7	0.12	8819.11	1.41	0.68	5.14	412.34	0	0.11	6.89
T8	0	-	1.12	0.59	5.14	671.73	0	0.18	7.12

by running the MIGP model without a time limit until the exact Pareto front was returned.

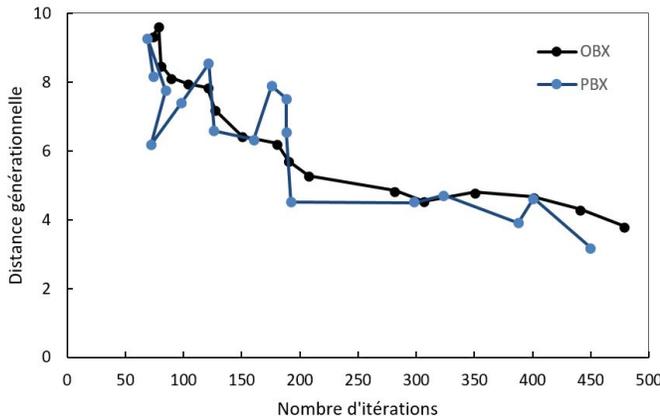


Fig. 4. Results of Crossover Operators

3) *Results of Quotas Calculation Procedure:* In this section, we present the experimental results of the MILP presented in Section V-A. Recall that, this mathematical model is used by the hybrid method to calculate an allocation (maximum quotas) of employees on projects. This allocation becomes the starting point for generating an approximate Pareto front using the NSGA-II method.

The MILP model was tested on the different sets of instances. Table VI shows the computational results obtained after a time limit of 5 minutes per instance. In this table, from left to right, the columns refer to the type of instance, the number of feasible instances, the number of solutions that are proven to be the optimal solutions, the average time required to prove optimality, and the average deviations from optimal for instances that are not optimally solved.

4) *NSGA-II and Greedy Heuristic Comparison:* The objective of the experiments presented in this section is to evaluate the improvement provided by the NSGA-II over the initial solutions obtained by the GH. To this end, we compare the results of both methods with those obtained by the MIGP. Note that, we use the best combination of NSGA-II parameters presented in Table V. The solutions returned by the GH are collected after a computation time (AG_{max}) limited to 5 minutes. We run both methods 10 times and the average values are calculated. For each run of the genetic algorithm, we used the same initial solutions generated by the GH, in order to ensure the consistency of the tests.

Table VII compares the experimental results of the NSGA-II and greedy heuristic with those obtained with the GPNE model. We used four different performance indicators: the average Pareto front size (PFS), the average hypervolume (HV), the average generational distance (GD), and the average computation time in seconds (in the CPU column).

5) *Experimental Analysis:* The comparison results in Fig. 4 show that the OBX operator has a better performance compared to the other PBX operator. In fact, the genetic algorithm has a better and faster convergence with the OBX genetic operator. Moreover, reassuringly it was found that computation time for both crossover operations is almost the same for each problem size.

The experiments in Table VI prove the performance of the MILP model in terms of feasibility and optimality. The model finds a feasible solution for each instance on a set. Moreover, all the solutions obtained for the instances on sets $T1, \dots, T4$ are optimal solutions. The model finds only 2, 5, 8 non-optimal solutions (among 40) for the sets of instances $T5, T6, T7$, respectively. For $T8$ instances, a large number of instances (11) are not solved to the optimum. However, the deviation from the optimum is very acceptable (8).

Finally, the results presented in Table VII allow us to clearly conclude that the NSGA-II algorithm significantly improves the solutions obtained by the greedy algorithm. All the metrics used show that the quality of the solutions is always better for all types of instances. Thus, we can also see that the computation time is very acceptable for a heuristic.

VII. CONCLUSION

We studied herein a two-agent multi-skill resource-constrained scheduling problem with a global objective. Motivated by a real scheduling case, we considered that each agent manages one or more projects and wants to minimize the total weighted tardiness of its tasks. We consider a pool of employees in which each one can perform a set of skills with heterogeneous performance levels. We assumed that there are some constraints that can be violated when there is no feasible schedule for the problem. Thus, the global objective function seeks to minimize the constraint violations by reducing the undesirable deviations in the soft constraints from their respective goals. The overall objective is to find a schedule that minimizes at the same time both agents objective functions and the global objective function. We provided a mixed-integer goal programming (MIGP) formulation for the problem. Furthermore, we provided a hybrid algorithm combining an exact procedure, a greedy heuristic, and a genetic algorithm to find

an approximate Pareto solution set. The performance of the heuristics is evaluated on a set of simulated instances. The results show that the NSGA-II algorithm is the best performing method.

Finally, in future research, we will also focus on other types of two-agent multi-skilled resources scheduling problems. Such as the constrained optimization problem, where the goal is to minimize the global function, subject to the constraints that the objective values of the other do not exceed a given threshold.

REFERENCES

- [1] A. Agnetis, P. B. Mirchandani, D. Pacciarelli, and A. Pacifici, "Scheduling problems with two competing agents," *Operations Research*, vol. 52, no. 2, pp. 229–242, 2004. [Online]. Available: <https://doi.org/10.1287/opre.1030.0092>
- [2] K. Baker and J.-C. Smith, "A multiple-criterion model for machine scheduling," *Journal of Scheduling*, vol. 6, pp. 7–16, 2003. [Online]. Available: <https://doi.org/10.1023/A:1022231419049>
- [3] P. Brucker, B. J. Bernd, and A. Krämer, "Complexity of scheduling problems with multi-purpose machines," *Annals of Operations Research*, vol. 70, no. 0, pp. 57–73, 1997.
- [4] E. Néron and D. Baptista, "Lower bounds for the multi-skill project scheduling problem," in *the Eighth International Workshop on Project Management and Scheduling*, 2002, pp. 274–277.
- [5] B. Afshar-Nadjafi, "Multi-skilling in scheduling problems: A review on models, methods and applications," *Computers and Industrial Engineering*, vol. 151, no. November 2020, p. 107004, 2021.
- [6] S. Hartmann and D. Briskorn, "An updated survey of variants and extensions of the resource-constrained project scheduling problem," *European Journal of Operational Research*, 2021.
- [7] M. C. Wu and S. H. Sun, "A project scheduling and staff assignment model considering learning effect," *The International Journal of Advanced Manufacturing Technology*, vol. 28, pp. 1190–1195, 2006.
- [8] C. Heimerl and R. Kolisch, "Scheduling and staffing multiple projects with a multi-skilled workforce," *OR Spectrum*, vol. 32, no. 2, pp. 19–25, 2010.
- [9] R. Kolisch and C. Heimerl, "An efficient metaheuristic for integrated scheduling and staffing it projects based on a generalized minimum cost flow network," *Naval Research Logistics (NRL)*, vol. 59, no. 2, pp. 111–127, 2012.
- [10] T. Felberbauer, W. J. Gutjahr, and K. F. Doerner, "Stochastic project management: multiple projects with multi-skilled human resources," *Journal of Scheduling*, vol. 22, pp. 271–288, 2019.
- [11] M. Walter and J. Zimmermann, "Minimizing average project team size given multi-skilled workers with heterogeneous skill levels," *Computers & Operations Research*, vol. 70, pp. 163–179, 2016.
- [12] W. J. Gutjahr, S. Katzensteiner, P. Reiter, C. Stummer, and M. Denk, "Multi-objective decision analysis for competence-oriented project portfolio selection," *European Journal of Operational Research*, vol. 205, no. 3, pp. 670–679, 2010.
- [13] R. Chen, C. Liang, D. Gu, and J. Leung, "A multi-objective model for multi-project scheduling and multi-skilled staff assignment for it product development considering competency evolution," *International Journal of Production Research*, vol. 55, no. 21, pp. 6207–6234, 2017.
- [14] M. Hematian, M. Esfahani, I. Mahdavi, N. Mahdavi-Amiri, and J. Rezaeian, "A multiobjective integrated multiproject scheduling and multi-skilled workforce assignment model considering learning effect under uncertainty," *Computational Intelligence*, vol. 36, no. 1, pp. 276–296, 2020.
- [15] M. Walter, *Multi-Project Management with a Multi-Skilled Workforce*. Springer Gabler, Wiesbaden, 2015.
- [16] L. Cui, X. Liu, S. Lu, and Z. Jia, "A variable neighborhood search approach for the resource-constrained multi-project collaborative scheduling problem," *Applied Soft Computing*, vol. 107, p. 107480, 2021.
- [17] C. T. Ng, T. C. Cheng, and J. J. Yuan, "A note on the complexity of the problem of two-agent scheduling on a single machine," *Journal of Combinatorial Optimization*, vol. 12, no. 4, pp. 386–393, 2006.
- [18] A. Agnetis, G. De Pascale, and D. Pacciarelli, "A lagrangian approach to single-machine scheduling problems with two competing agents," *Journal of Scheduling*, vol. 12, no. 4, pp. 401–415, 2009.
- [19] J. Li, Y. Gajpal, and S. S. Appadoo, "Algorithms for a two-agent single machine scheduling problem to minimize weighted number of tardy jobs," *Journal of Information and Optimization Sciences*, vol. 42, no. 4, pp. 785–811, 2021.
- [20] X. Zhang, "Two competitive agents to minimize the weighted total late work and the total completion time," *Applied Mathematics and Computation*, vol. 406, p. 126286, 2021.
- [21] W. C. Lee, J. Y. Wang, and H. W. Su, "Algorithms for single-machine scheduling to minimize the total tardiness with learning effects and two competing agents," *Concurrent Engineering Research and Applications*, vol. 23, no. 1, pp. 13–26, 2015.
- [22] T. C.E.Cheng, S.-R. Cheng, W.-H. Wu, P.-H. Hsu, and C.-C. Wu, "A two-agent single-machine scheduling problem with truncated sum-of-processing-times-based learning considerations," *Computers and Industrial Engineering*, vol. 60, no. 4, pp. 534–541, 2011. [Online]. Available: <http://dx.doi.org/10.1016/j.cie.2010.12.008>
- [23] Y. Yin, S.-R. Cheng, T. Cheng, C.-C. Wu, and W.-H. Wu, "Two-agent single-machine scheduling with assignable due dates," *Applied Mathematics and Computation*, vol. 219, no. 4, pp. 1674–1685, 2012.
- [24] D.-J. Wang, Y. Yin, J. Xu, W.-H. Wu, S.-R. Cheng, and C.-C. Wu, "Some due date determination scheduling problems with two agents on a single machine," *International Journal of Production Economics*, vol. 168, pp. 81–90, 2015.
- [25] E. Gerstl and G. Mosheiov, "Scheduling problems with two competing agents to minimized weighted earliness–tardiness," *Computers & Operations Research*, vol. 40, no. 1, pp. 109–116, 2013. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0305054812001359>
- [26] S. S. Li, R. X. Chen, and Q. Feng, "Scheduling two job families on a single machine with two competitive agents," *Journal of Combinatorial Optimization*, vol. 32, no. 3, pp. 784–799, 2016.
- [27] S. N. Sahu, Y. Gajpal, and S. Debbarma, "Two-agent-based single-machine scheduling with switchover time to minimize total weighted completion time and makespan objectives," *Annals of Operations Research*, vol. 269, no. 1-2, pp. 623–640, 2018.
- [28] B.-C. Choi, M.-J. Park, and J. Du, "Scheduling two projects with controllable processing times in a single-machine environment," *Journal of Scheduling*, vol. 23, no. 1, pp. 619–628, 2020.
- [29] H. Balasubramanian, J. Fowler, A. Keha, and M. Pfund, "Scheduling interfering job sets on parallel machines," *European Journal of Operational Research*, vol. 199, no. 1, pp. 55–67, 2009.
- [30] K. Zhao and X. Lu, "Approximation schemes for two-agent scheduling on parallel machines," *Theoretical Computer Science*, vol. 468, pp. 114–121, 2013.
- [31] W.-C. Lee, J.-Y. Wang, and M.-C. Lin, "A branch-and-bound algorithm for minimizing the total weighted completion time on parallel identical machines with two competing agents," *Knowledge-Based Systems*, vol. 105, pp. 68–82, 2016.
- [32] K. Zhao and X. Lu, "Two approximation algorithms for two-agent scheduling on parallel machines to minimize makespan," *Journal of Combinatorial Optimization*, vol. 31, no. 1, pp. 260–278, 2016. [Online]. Available: <http://dx.doi.org/10.1007/s10878-014-9744-y>
- [33] J. Y.-T. Leung, M. Pinedo, and G. Wan, "Competitive two-agent scheduling and its applications," *Operations Research*, vol. 58, no. 2, pp. 458–469, 2010.
- [34] G. Wan, S. R. Vakati, J. Y.-T. Leung, and M. Pinedo, "Scheduling two agents with controllable processing times," *European Journal of Operational Research*, vol. 205, no. 3, pp. 528–539, 2010.
- [35] D. Li and X. Lu, "Two-agent parallel-machine scheduling with rejection," *Theoretical Computer Science*, vol. 703, pp. 66–75, 2017. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0304397517306527>
- [36] J. Pei, J. Wei, B. Liao, X. Liu, and P. M. Pardalos, "Two-agent scheduling on bounded parallel-batching machines with an aging effect of job-position-dependent," *Annals of Operations Research*, vol. 294, no. 1-2, pp. 191–223, 2020.

- [37] F. Sadi and A. Soukhal, "Complexity analyses for multi-agent scheduling problems with a global agent and equal length jobs," *Discrete Optimization*, vol. 23, pp. 93–104, 2017.
- [38] D. Elvikis, H. W. Hamacher, and V. T'kindt, "Scheduling two agents on uniform parallel machines with makespan and cost functions," *Journal of Scheduling*, vol. 14, no. 5, pp. 471–481, 2011.
- [39] D. Elvikis and V. T'kindt, "Two-agent scheduling on uniform parallel machines with min-max criteria," *Annals of Operations Research*, vol. 213, no. 1, pp. 79–94, 2014.
- [40] Y. Yin, S.-R. Cheng, T. Cheng, D.-J. Wang, and C.-C. Wu, "Just-in-time scheduling with two competing agents on unrelated parallel machines," *Omega*, vol. 63, pp. 41–47, 2016. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0305048315002042>
- [41] Y. Yin, Y. Chen, K. Qin, and D. Wang, "Two-agent scheduling on unrelated parallel machines with total completion time and weighted number of tardy jobs criteria," *Journal of Scheduling*, vol. 22, no. 3, pp. 315–333, 2019.
- [42] R. Graham, E. Lawler, J. Lenstra, and A. Kan, "Optimization and approximation in deterministic sequencing and scheduling: a survey," in *Discrete Optimization II*, ser. Annals of Discrete Mathematics, P. Hammer, E. Johnson, and B. Korte, Eds. Elsevier, 1979, vol. 5, pp. 287–326.
- [43] A. Agnetis, J.-C. Billaut, S. Gawiejnowicz, D. Pacciarelli, and A. Soukhal, *Multiagent Scheduling: Models and Algorithms*. Springer, Berlin, Heidelberg, 2014.
- [44] K. Deb, A. Pratap, S. Agarwal, and T. Meyarivan, "A fast and elitist multiobjective genetic algorithm: NSGA-II," *IEEE Transactions on Evolutionary Computation*, vol. 6, no. 2, pp. 182–197, 2002.
- [45] N. Srinivas and K. Deb, "Multiobjective Optimization Using Non-dominated Sorting in Genetic Algorithms," *Evolutionary Computation*, vol. 2, no. 3, pp. 221–248, 1994.
- [46] G. Syswerda, *Scheduling optimization using genetic algorithms*, New York, NY, 1991.
- [47] T. Kellegöz, B. Toklu, and J. Wilson, "Comparing efficiencies of genetic crossover operators for one machine total weighted tardiness problem," *Applied Mathematics and Computation*, vol. 199, no. 2, pp. 590–598, 2008.
- [48] E. Zitzler and L. Thiele, "Multiobjective optimization using evolutionary algorithms — a comparative case study," in *Parallel Problem Solving from Nature — PPSN V*, A. E. Eiben, T. Bäck, M. Schoenauer, and H.-P. Schwefel, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 1998, pp. 292–301.
- [49] D. A. V. Veldhuizen and G. B. Lamont, "Multiobjective evolutionary algorithms: Analyzing the state-of-the-art," *Evolutionary Computation*, vol. 8, no. 2, pp. 125–147, 2000.

Transformer based Model for Coherence Evaluation of Scientific Abstracts: Second Fine-tuned BERT

Anyelo-Carlos Gutierrez-Choque¹, Vivian Medina-Mamani², Eveling Castro-Gutierrez³,
Rosa Núñez-Pacheco⁴, Ignacio Aguaded⁵
Universidad Nacional de San Agustín de Arequipa, Perú^{1,2,3,4}
Universidad de Huelva, España.⁵

Abstract—Coherence evaluation is a problem related to the area of natural language processing whose complexity lies mainly in the analysis of the semantics and context of the words in the text. Fortunately, the Bidirectional Encoder Representation from Transformers (BERT) architecture can capture the aforementioned variables and represent them as embeddings to perform Fine-tunings. The present study proposes a Second Fine-Tuned model based on BERT to detect inconsistent sentences (coherence evaluation) in scientific abstracts written in English/Spanish. For this purpose, 2 formal methods for the generation of inconsistent abstracts have been proposed: Random Manipulation (RM) and K-means Random Manipulation (KRM). Six experiments were performed; showing that performing Second Fine-Tuned improves the detection of inconsistent sentences with an accuracy of 71%. This happens even if the new retraining data are of different language or different domain. It was also shown that using several methods for generating inconsistent abstracts and mixing them when performing Second Fine-Tuned does not provide better results than using a single technique.

Keywords—Coherence evaluation; inconsistent sentences detection; BERT; second fine-tuned

I. INTRODUCTION

Natural language processing (NLP) is a subarea of artificial intelligence that involves tasks related to the analysis of text information using computational means. These tasks are: text generation, automatic text summarization, speech analysis and information extraction. Textual coherence modeling belongs to this class of tasks described; it consists of distinguishing coherent documents from incoherent ones [1]. Coherence in NLP is very relevant nowadays, because it is implicitly involved in several applications such as speech generation, text summarization generation, translations etc. The models proposed in text generation must ensure that the results are coherent texts. The automatic evaluation of coherence contributes to the generation of these texts with quality.

According to Charolles [2], coherence operates through the thematic progression which implies that all the ideas of a coherent text must be connected to each other. Each sentence provides a piece of information that ensures thematic continuity. A coherent text must also present consistency of ideas; this implies that no idea in a text should contradict another and neither should it be incongruent with the universe of the text to which it belongs.

Coherence also implies the type of informational and semantic connectivity that a text possesses [3]. A text is

considered coherent if it is semantically consistent and provides cognitive integrity [4], therefore, a coherent document is easier to understand than an incoherent document. Coherence is more important when analyzing scientific papers, as it must communicate information effectively to reviewers and researchers. Incoherence in scientific writing directly affects both the reading experience and the comprehensibility of scientific papers [27]. Let us consider the sentence-divided scientific papers in Fig. 1 and 2.

In the left column of Fig. 1, the scientific abstract reports on the effects of candy advertising and consumption reactions of certain additives in children under 12 years old, while in the right column, the third sentence reports on project-based learning that expresses an idea different from the other sentences, evidencing the incoherence of the text, due to the fact that the thematic progression and consistency of ideas have not been met. This phenomenon occurs in the same way in the right column of Fig. 2. The abstract deals with machine learning and data representation, while the sixth sentence reports on image processing.

As we have seen, the incoherence that occurs during scientific writing creates difficulties in transmitting and disseminating the authors ideas [27]. This happens because the sentences produced are not strongly interconnected, but are isolated, managing to label the scientific paper as "poorly written" or "difficult to follow" [28]. This kind of problem can be intentional or unintentional. Unfortunately, most of the existing systems that check for errors in scientific papers [29] lack advanced features for coherence quality control.

Given the above consideration, identifying incoherent sentences in a scientific paper becomes a problem of high rigor when evaluating scientific abstracts. The abstract is the only part of the article that is usually published in conference proceedings, and that readers usually review when searching through electronic databases; Likewise, it is a section that a potential referee gives a reading to when they are invited by an editor to review a manuscript [5]. It often contains the following structure: context, methodology, results and conclusions, each of which should provide relevant and semantically consistent information.

The evaluation of coherence requires a thorough analysis of the parts of the text at the structural and semantic level, since being a natural language it does not follow a set of rules like formal languages [4]. It is too abstract a concept [6]; however, it is a problem that has been given attention in different studies

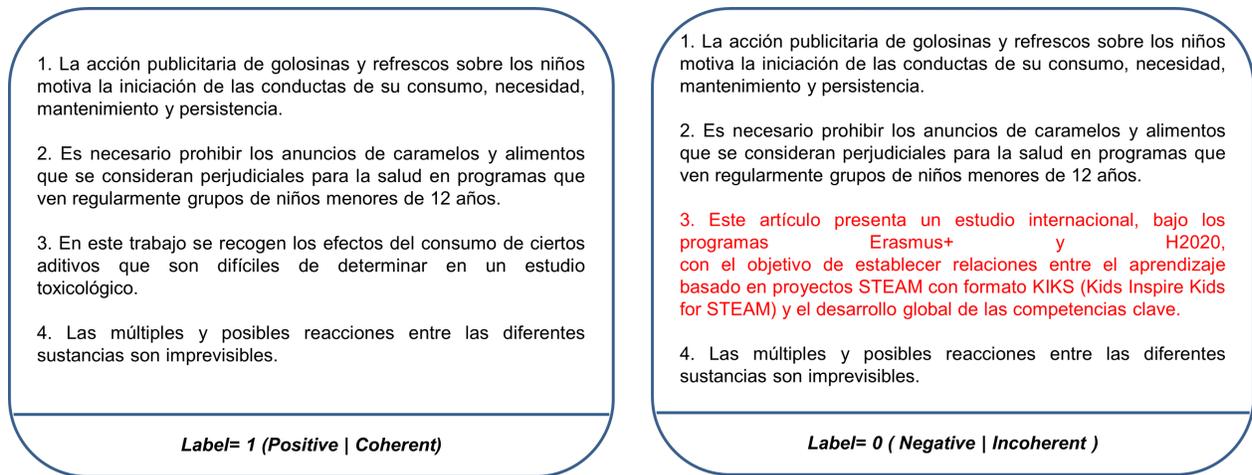


Fig. 1. Coherent and Incoherent Scientific Abstract in Spanish.

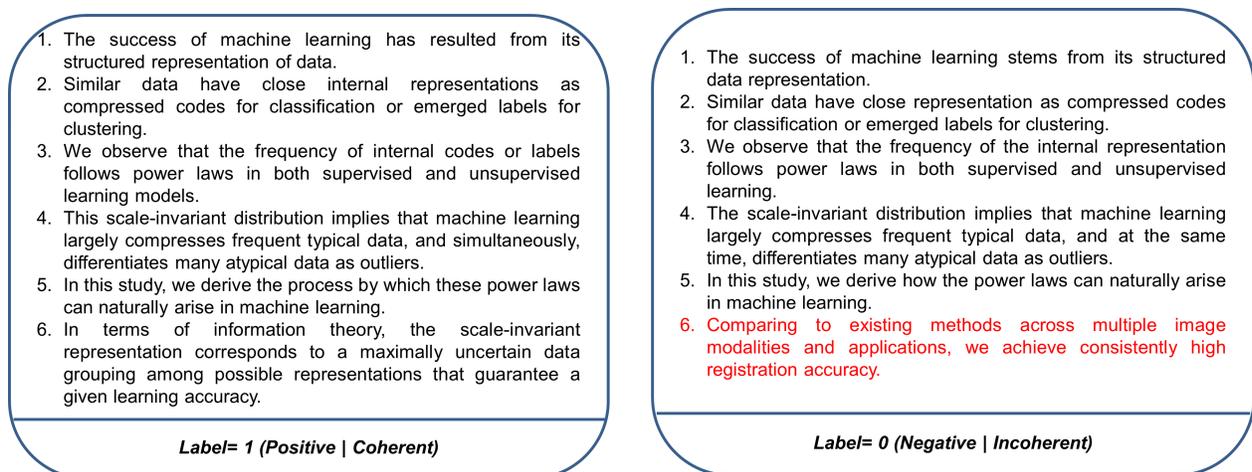


Fig. 2. Coherent and Incoherent Scientific Abstract in English.

and/or solutions since the twentieth century, as mentioned in the following paragraphs.

Foltz in 1998 [7] proposed the first coherence evaluation method using machine. This method is based on the study of latent semantic analysis (LSA), a method that compares units of textual information and determines their semantic relationship. In the following years, several coherence analysis methods were proposed by various researchers; however, no method has proved to be perfect [8].

Relational models such as the "Rhetorical Structure Theory" [9] define relationships that hierarchically structure texts. Thus, in the work of Barzila and Lapata (2008) a model called Entity Grid [10] was proposed to evaluate local cohesion. This approach was based on the centering theory [11], which models a text as a set of segments and utterances that produce centers of attention.

Unlike the Entity Grid model [10], which is a method for evaluating coherence at the local level, in the work of Guinaudeau and Strube (2013) a graphical model called Entity Graph [12] was proposed to measure text coherence at the global level. This bipartite graph allows relating non-adjacent

sentences of a text.

In the work of Li and Hovy (2014) it has been shown that recurrent and recursive neural networks are designed to estimate the coherence of a text [13]. Recurrent neural networks simulate the processing of a text according to a reading process: word by word; while in the recursive neural network the processing is represented through a binary tree.

Li and Jurafsky (2016) developed a discriminative neural model that can distinguish coherent and incoherent text. They also created 2 generative models that produce coherent text, one is based on SEQ2SEQ and the other is a Markovian model. These models capture the latent discourse dependencies of a given text [14]

Based on the foundations of the Entity Graph model [12], a semantic similarity graph model was proposed in the work of Putra and Tokunaga (2017) to address coherence from a cohesion perspective [15]. They argue that the coherence of a text is built by the cohesion between its sentences. This method employs an unsupervised learning approach.

In the work of Baiyun Cui *et al.* (2017) a deep coherence

model (DCM) was proposed making use of a convolutional neural network architecture to capture text coherence [6]. The model captures the interactions between sentences by calculating the similarities of their distributional representations.

In the work of Mesgar and Strube [21], a local coherence model was developed using a unidirectional standardized LSTM architecture to encode the context of an input sequence of words, then the relationships between adjacent sentences were encoded using LSTM. Finally, a vector representing the coherence of the text was produced.

The use of Recurrent or Recursive Neural Networks allows a vector representation of an input sequence. Based on this, a series of dense (linear) layers can be applied to classify whether the input sequence is coherent or incoherent. Thus, in the work of Moon *et al.* [22], the Bidirectional LSTM (BiLSTM) sentence encoder was applied to capture the grammar of each sentence. Given the numerical representations of the sentences, the local coherence model and the global coherence model extract the respective features.

Bao *et al.* [23] used Recurrent Neural Networks (RNN) for the model to semantically represent a text. For this purpose, they used bidirectional closed recurrent units (BiGRU) in conjunction with the pretrained language model Word2Vec to represent this semantics. The results show that a complete analysis of the coherence of a text can be represented, which favors the task of binary text classification.

The creation of a dataset for training, validation and testing is also indispensable for the evaluation of coherence. Because of this, the work of Mohammadi *et al.* [24] proposes different techniques for generating incoherent or negative documents to be added to the coherent documents in order to train a Convolutional Neural Network. Their results indicate that artificially generating incoherent documents does not guarantee "sufficiently incoherent" documents, which negatively influences the accuracy of the model.

As seen, RNN, LSTM, gated recurrent neural network (GRNN) and BiLSTM are some of the sequence models for NLP tasks such as natural language modeling [17]. In 2017, the Google research team presented the Transformer architecture that replaces the complex RNN and CNN (Convolutional Neural Network) architectures because of its better results: parallel training capability with several GPUs and self-attention mechanism, which allows to "remember" the information in the long term [18]. Fig. 3 shows the architecture of the Transformer and the Bidirectional Encoder Representations from Transformers model. In the Transformer architecture, the encoder maps an input sequence of symbols $(x_1, x_2, \dots, x_{n-1}, x_n)$ to a sequence of continuous representations $z = (z_1, z_2, \dots, z_{n-1}, z_n)$. Given z , the decoder generates a sequence of output symbols $(y_1, y_2, \dots, y_{m-1}, y_m)$, considering one element at a time. The general architecture of Transformer comprises the self-attention mechanism, the encoder and decoder, which are fully connected.

BERT is a model of the pretrained open source language introduced in 2018 [19]. It is based on Google's Transformer architecture. Also, it is designed to pre-train text representations in a bidirectional (left-to-right) manner from unlabeled texts [20]. BERT has two pre-trained models: BERT Base and BERT Large. The first model consists of 12 encoders

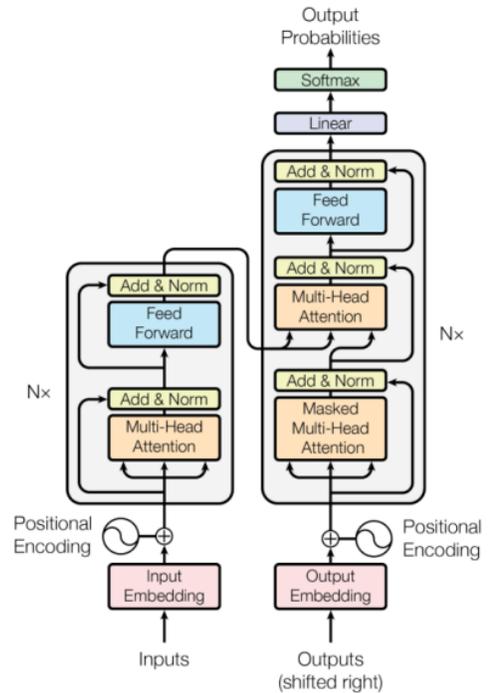


Fig. 3. The Transformer Model Architecture [18].

and a bidirectional self-attention mechanism; while the second model consists of 24 encoders and 16 bidirectional heads. The BERT model is pre-trained with 800 million words from BooksCorpus and unlabeled text from English Wikipedia with 2.5 billion words. This model is suited for small datasets related to specific NLP tasks, for example, the evaluation of coherence in scientific papers.

Fig. 4 shows the neural network architecture of the deep bidirectional BERT and unidirectional (from left to right) OpenAI GPT contextual models [17], in which the unidirectional model generates a representation for each word based on other words in the same sentence. The bidirectional BERT model represents both the preceding and following context in a sentence. However, the context-free models Word2vec and Glove generate a word representation based on each vocabulary word.

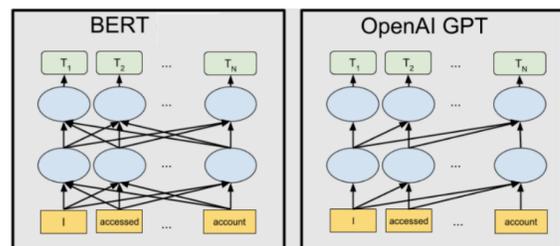


Fig. 4. BERT and OpenAI GPT Neural Network Architecture [18].

According to the historical review, BERT has revolutionized the field of NLP by enabling transfer learning of large language models that can capture complex textual patterns [36]. It also has the advantage of offering better performance and scalability over recurrent neural network architectures

since the latter operate sequentially, while BERT can be parallelized. This research work proposes a Second Fine-Tuned model based on BERT to detect inconsistent sentences to evaluate the coherence in abstracts written in Spanish/English. Two formal techniques for generating incoherent abstracts are also proposed in order to improve the training and validation of the aforementioned model.

This paper is organized as follows: in Section II the most recent and important research on the evaluation of coherence is mentioned. In Section III the methodology and the research proposal are described. Section IV presents the experiment with the results. Finally, Section V presents the discussion, conclusions and future work.

II. RELATED WORK

Discovering semantic progression and consistency of ideas is indispensable for understanding coherence. Previous research has relied on the RNN, LSTM and BiLSTM architecture to evaluate coherence; however, these networks do not use a self-attention mechanism to encode sentences and some information is lost [16]. The Transformer-based architecture allows receiving input sequences in parallel making it more efficient; and, specifically, BERT allows capturing the context of a sentence based on bidirectional analysis [18]. Some work based on BERT to evaluate coherence is shown below.

In the work of Muangkammuen *et al.* (2020), a scoring method based on BERT was proposed to score text clarity using local coherence between adjacent sentences. Cause-effect relationship and contrast were considered [25]. First, a local coherence model was trained according to BERT; then the model was retrained to evaluate the clarity of a text. The results show that retraining provides positive results even if the data on which both trainings were performed were not domain related.

In the work of Callan and Foster (2021), a corpus of narrative stories automatically generated by the pre-trained Transformer GPT-NEO model was proposed, which were analyzed by humans and by 2 automated metrics: BERT Score y BERT NSP. This was done in order to evaluate the coherence and level of interest of narrative texts. The results show that greater emphasis should be placed on BERT evaluation techniques and that generative models do not always produce coherent texts; the more natural and coherent the generated text is, the higher its quality [26].

In the work of *et al.* (2021), a comparative analysis of 3 different types of models for the evaluation of coherence in Polish documents has been developed. The first one is based on Semantic Similarity Graph (SSG); the second one is based on Long Short Term Memory (LSTM); and the third one is based on BERT. The results show that the neural network related methods offer better accuracy than the SSG related methods; and within the neural networks, although the LSTM based method shows better accuracy compared to the BERT based method, it is emphasized that the latter can increase the value of said metric with an additional Fine-Tuned [4].

In the work of Nguyen and Zaslavskiy (2021), a Fine-Tuned method based on the BERT model in conjunction with a clustering algorithm was proposed for the detection

of discordant sentences in a corpus of scientific documents written in English and Russian, in order to detect incoherence in scientific writing. Primero generaron ejemplos negativos mediante la métrica BERT Score para calcular la similitud semántica entre pares de oraciones. They first generated negative examples using the BERT Score metric to compute semantic similarity between sentence pairs. Then they trained a model with coherent and incoherent sentence pairs. Finally, they retrained this model with whole paragraph training. The results were positive [27].

In the work of Bendeviski *et al.* (2021), a comparative analysis of different artificial intelligence methods for predicting the coherence score of narrative documents was proposed, where it was evidenced that BERT produces better results compared to traditional machine learning methods such as: Linear Regression, Support Vector Machine, Random Forest. They establish dimensions for coherence which are: Context, Chronology and Theme, each of these dimensions possess narrative texts with a coherence score of 0-3 (4-class classification) [30].

According to the work of Noji and Takamura [31], negative examples contribute to a neural language model's ability to robustly handle complex syntactic constructs and improve its robustness.

III. PROPOSED MODEL AND METHODOLOGY

The principal objective of this study is to build a Second Fine-Tuned model based on BERT to detect inconsistent sentences of abstracts in English/Spanish (coherence evaluation). Two negative example generation techniques have been used for this coherent scientific abstracts are positive examples ($label = 1$), while incoherent scientific abstracts are negative examples ($label = 0$).

A. Data Recollection and Preprocessing

First of all, a program has been developed in Python with the beautifulsoup4 library to perform web scraping to the website of the journal "Comunicar" [32]. From this journal 1,493 scientific abstracts written in Spanish were extracted. Second, the corpus of "Medical Semantic Indexing in Spanish" (MESINESP) was downloaded [33], from which 51,390 scientific abstracts written in Spanish were collected. Thirdly, a corpus of arXiv was downloaded through Kaggle [34], from which 56,181 scientific documents written in English were collected.

In addition, a corpus of 448 scientific abstracts from the "International Conference on Machine Learning and Applications" (ICMLA) were added to this corpus. [35]. As a result, a corpus of 56,629 scientific documents was constructed, 3 corpus were collected, 2 in Spanish and 1 in English. It should be added that the downloaded corpus were subjected to 2 preprocessing stages: Remove blanks and Remove duplicates. Two formal techniques were developed for the generation of negative examples (incoherent abstracts):

1) *Random Manipulation (RM)*: In the first method, every abstract T_i of a corpus D is tokenized in N sentences. This T_i is represented as a set of sentences $T_i = \{S_1, S_2, S_3, \dots, S_{N1}, S_N\}$. Every S_j is a sentence where j is

the position of the sentence in T_j , it must be fulfilled that $1 \leq j \leq N$. The variable i represents the position of a scientific abstract in the corpus D , it must be fulfilled that $1 \leq i \leq size(D)$. Then a S_j is randomly selected, knowing that: $S_j \in T_i, T_i \in D$. Therefore, this S_j is replaced with a $S_{j'}$; knowing that: $S_{j'} \in T_{i'}, T_{i'} \in D$, and $S_j \neq S_{j'}$.

2) *Manipulation Random based K-means (KRM)*: In the second method, embeddings with BERT are generated from all abstracts of a corpus D . Then by clustering with K-Means ($\#clusters : K = 10$) As a partial result, there are 10 clusters of scientific abstracts labeled as C_i , being i the cluster number in the corpus D . These clusters are grouped by similar embeddings. The K-means algorithm minimizes the principle of inertia, according to the following equation 1:

$$\sum_{i=0}^n \min_{\mu_j \in C} (\|x_i - \mu_j\|^2) \quad (1)$$

This measure 1 indicates the internal coherence between the samples belonging to the different clusters. Taking as a reference the variables of the first method; to generate negative examples a $S_k \in T_j$ is randomly chosen, considering that: $T_j \in C_i$ y $C_i \in D$. Then this S_k is randomly replaced with a $S_{k'}$ knowing that: $S_{k'} \in T_{j'}, T_{j'} \in C_{i'} \text{ y } C_{i'} \in D$. It must be fulfilled that: $S_k \neq S_{k'}$ y $C_i \neq C_{j'}$. Having clusters whose similar scientific documents, knowing that the groups among themselves are different, it is ensured that negative examples are explicitly more incoherent than the first RM method. Finally, the corpus are summarized in the following table I.

TABLE I. CORPUS SUMMARY

Features	Comunicar Corpus	MESINESP Corpus	arXiv + ICMLA Corpus
Language	Spanish	Spanish	English
Coherent abstracts	1,493	50,047	56,181
Incoherent abstracts	RM Method	RM Method	RM Method
	1,491	1,491	41,559
Total abstracts	2,984	2,984	97,740

B. Proposed BERT Model

A Second Fine-Tuned model based on BERT is proposed to detect inconsistent sentences in scientific abstracts written in English/Spanish (coherence evaluation), six different experiments have been carried out for this purpose, the same ones as detailed below:

- 1) First Fine-Tuned to the original pre-trained model of BERT with the Spanish-language dataset of the journal "Comunicar". The dataset has been divided into three segments: training, validation and testing. The technique used to generate negative examples was RM.
- 2) Second Fine-Tuned to the model trained in experiment 1 by mixing the MESINESP Spanish dataset and "Comunicar" dataset, the same testing set as experiment 1 has been maintained. The technique used to generate negative examples was RM.

- 3) Second Fine-Tuned to the model trained in experiment 1 by mixing the dataset in English of "arXiv + ICMLA" with that of "Communicate". The same testing segment has been maintained as experiment 1. The technique used to generate negative examples was RM.
- 4) First Fine-Tuned to the original pre-trained model of BERT with the Spanish-language dataset of the journal "Comunicar". The dataset has been divided into 2 segments: training and validation. The same testing segment has been maintained as experiment 1. The technique used to generate negative examples was KRM.
- 5) Second Fine-Tuned to the model trained in experiment 4. The same data segments of experiment 2 and also the same testing segment of experiment 1 have been maintained.
- 6) Second Fine-Tuned to the model trained in experiment 4. The same data segments of experiment 3 and also the same testing segment of experiment 1 have been maintained.

The purpose of the experiments described above is to determine which model is best for evaluating coherence in scientific abstracts written in English/Spanish. Once the positive and negative examples (datasets) have been generated, these experiments have followed the framework proposed in Fig. 5

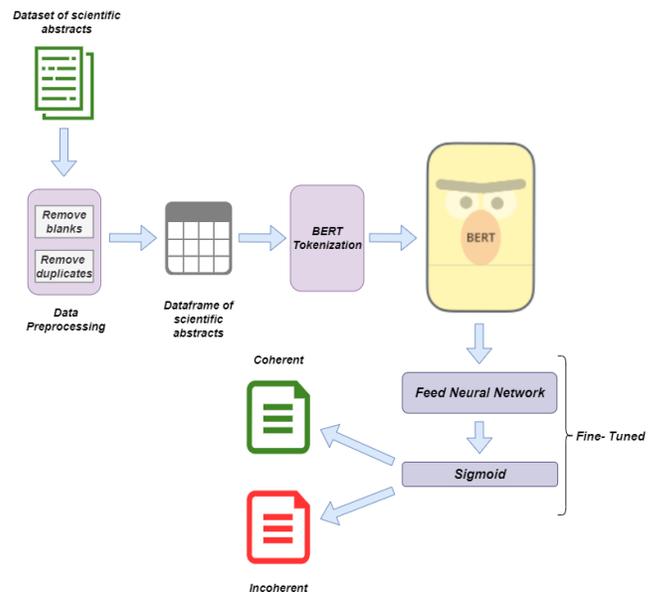


Fig. 5. BERT-Large Coherence Evaluation Framework.

During the tokenization process, an input sequence must be preprocessed to produce tokens (prayer units). In this process, the CASED preprocessor BERT was used, which is a multilingual preprocessor that can process special characters or capital letters/minuscul. A sorting token [CLS] is always included at the beginning of a sentence, each word being a token. To separate one sentence from another, the separation token [SEP] is included.[37]. It is known that an abstract T_i is represented as $T_i = \{S_1, S_2, S_3, \dots, S_{N1}, S_N\}$, applying the preprocessing you get: $\{S'_1, S'_2, \dots\}$.

The BERT Large pre-trained model was used to perform the numerical coding process of input sequences. This is a more complete version with 24 encoders and 340 million parameters [37]. It was ensured that this model used is multi-lingual and also that it is CASED. For each token 3 representations of embeddings were applied, the first is called token embeddings (h_{s1}), is responsible for representing a token as a numerical vector. The second is called segment embeddings (h_{s2}), this indicates to which segment a token embeddings belongs. It is known that a segment is that delimited by the [SEP] separator of another segment. The third is called positional embeddings (h_{s3}), this indicates the relative position of a token embeddings in the sentence. Each word is processed simultaneously [37]. Finally, each numerical representation is added to produce a single resulting vector h_c that will be used to train the Fine-Tuned model. The vector h_c can be represented by the following equation:

$$h_c = [h_{s1}, h_{s2}, h_{s3}] \quad (2)$$

Also, Fig. 6 represents how these embedding layers work with a pre-processed input. Fig. 7 shows the general architecture of the BERT Large model.

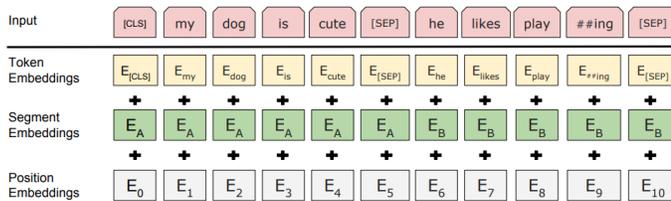


Fig. 6. BERT Input Representation [37].

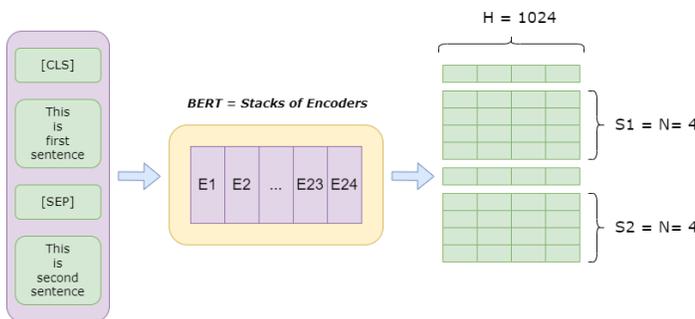


Fig. 7. Architecture BERT LARGE Encoders.

To perform the Fine-Tuned of the pre-trained model, following the six experiments described above, the dataset of the journal "Communicar" for testing has been divided by 10%. Then 10% was divided for validation and 80% for training. Each experiment has the same percentage of data division. It should be remembered that each experiment was tested with the testing data of the journal "Communicar". After defining the data sets, a Neural Network Layer Dense has been built with a Rectified Linear Unit activation function (ReLU). This activation function will generate a positive or zero output (if the input is negative). This function is optimal to accelerate

the training of deep neural networks, it is defined with the following equation 3:

$$f(x) = \max(0, x), x = input \quad (3)$$

After the ReLU layer, a Dropout layer was added to avoid overfitting the model during training; as several studies have shown that it reduces this overfitting and improves the performance of deep neural networks for tasks such as document classification [38]. After implementing the Neural Network layer, a simple layer has been added with a sigmoid activation function. This function is commonly used for binary classifications. It has a prediction range of [0 – 1], with those closest to zero being those incoherent abstracts (rounded to 0) and those closest to 1 being the coherent ones (rounded to 1). The mathematical representation of this function is shown in the following equation 4:

$$S(x) = \frac{1}{1 + e^{-x}}, x = input \quad (4)$$

The latter simple layer contains a single neuron, whose output represents the probability of coherence of a group of abstract sentences. It can be mathematically defined by the following equations 5 and 6. Each experiment has followed the same procedure so far mentioned. Considering the above, the 6 experiments possess the same architecture. This architecture is observed in the following Fig. 8.

$$q_c = f(W_{sen}^T h_c + b_{sen}) \quad (5)$$

$$output = \text{sigmoid}(U^T q_c + b) \quad (6)$$

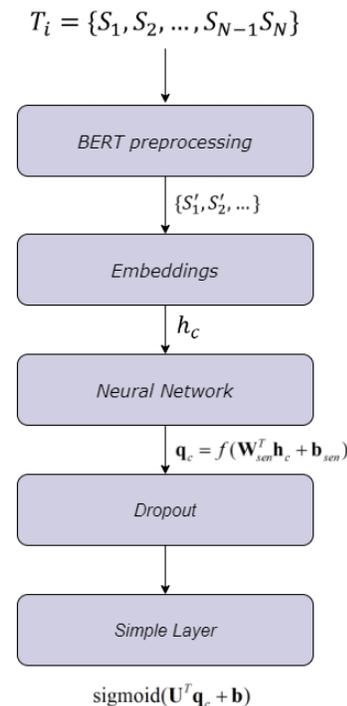


Fig. 8. Based BERT Model Architecture.

IV. EXPERIMENTAL SETTINGS AND RESULTS

In this section, the 6 experiments mentioned in the previous section are in depth. Six models with exact architecture have been generated to Fig. 8. The only difference between the experiments are the positive and negative data sets with which their models have been trained. The following subsections are detailed: The Datasets, Experimental Environments, Parameters Fine-tuning and Performance measurements:

A. The Datasets

According to Table I and described in previous sections; each experiment was assigned training, validation and testing data segments. This is shown in Table II. These datasets feed each model to perform the corresponding Fine-Tuned. It should be mentioned that each dataset generated for the experiments has passed through a preprocessing stage explained in Section III.

TABLE II. PREPARED DATASET

	Training (80%)	Validation (10%)	Testing (10%)	Dataset	Method
Experiment 1	2,416	269	299	Comunicar	RM
Experiment 2	90,000 + 2,416	10,000 + 269	299	MESINESP + Comunicar	RM
Experiment 3	87,966 + 2,416	9774 + 269	299	arXiv + ICMLA + Comunicar	RM
Experiment 4	2416	269	299	Comunicar	KRM
Experiment 5	90,000 + 2,416	10,000 + 269	299	MESINESP + Comunicar	RM + KRM
Experiment 6	87,966 + 2,416	9774 + 269	299	arXiv + ICMLA + Comunicar	RM + KRM

According to Table II, 2 formal methods of generating negative examples have been applied as explained in Section III. Experiment 1 uses only the dataset of the journal "Comunicar". Experiments 2 and 3 depend on experiment 1, in that sense; experiment 2 first uses the dataset of MESINESP and, secondly, the data of the journal "Comunicar" is added (repeating exactly the same training and validation data). Experiment 3 follows the same flow of experiment 2, only that it uses its own dataset as detailed in Table II. Incoherent abstracts of experiments 1,2 and 3 were generated with the RM method.

Experiment 4 is similar to experiment 1 with the difference that its incoherent abstracts were generated by the KRM method. Experiments 5 and 6 depend on experiment 4. In that sense, experiment 5 is similar to experiment 2, with the difference that when performing the training, the techniques of RM and KRM are combined to create a varied dataset. The RM method was applied to the MESINESP dataset and the KRM to the "Comunicar" dataset. Experiment 6 follows the same flow as experiment 5 with the difference that uses its own dataset as detailed in Table II.

In summary, the incoherent abstracts of experiment 4 were generated with the KRM method, therefore, in experiments 5 and 6 they ended up using both KRM and RM methods. It should be remembered that the testing data was generated with the RM method. This set belongs only to the magazine "Comunicar", repeating in each of the experiments without exception.

B. Experimental Environments

Experiments 1 and 4 were executed in the Google Colab environment. This environment serves to create automatic machine learning models for free and with powerful hardware resources such as: Graphic Processing Unit (GPU) and Tension Processing Unit (TPU) [39]. On the other hand, experiments 2, 3, 5 and 6 were executed on a standard server. Resources used for experiments are described in Table III.

TABLE III. EXPERIMENTAL COMPONENTS AND ENVIRONMENTS

Components	Details	
Dataset prepared	Libraries: Pandas, Beautiful Soup 4	
Preprocessor	bert_multicased_preprocess of Tensorflow.	
Model	bert_multi_cased_L-12_H-768_A-12 of Tensorflow.	
Language Programming & Tools	Python 3.9.6, Jupyter Notebook, Colab Notebook.	
Libraries & Frameworks	Tensorflow, Keras, Numpy, Pandas, Scikit-learn, Matplotlib, Seaborn, Nltk.	
Server	Environment 1 (Experiments 1,4)	Environment 2 (Experiments 2,3,5,6)
	Google Colab Cloud, GPU Tesla K80 11 GB Memory.	Ubuntu 64-bit S.O., Intel(R) Xeon(R) Gold 5115, CPU @ 2.40GHz, GPU Quadro P5000 16 GB Memory.

C. Parameters Fine-Tuning

The parameters, hyper-parameters and other configurations of the 6 experiments are detailed in the following Table IV.

TABLE IV. PARAMETERS SETTINGS

Name	Parameters and Hiper-Parameters	Value
Experiments 1, 2, 3, 4, 5, 6	- Model	- bert_multi_cased
	- Platform	- Tensorflow
	- Activation Function	- ReLU and Sigmoid
	- Dropout rate	- 0.1
	- Class_weight	- Balanced
	- Callback Model CheckPoint	- Max val_accuracy
	- Optimizer	- Adam
	- Learning rate	- 1e-05
	- Loss Function	- Binary Cross Entropy
	- Epochs	- 5
	- BatchSize	- 64

D. Performance Measures

The basic components that have been used for the evaluation of the models developed during the six experiments are the following:

- True Positive (TP): When the model correctly predicts the positive class. This means it correctly predicts a coherent abstract.
- True Negative (TN): When the model correctly predicts the negative class. This means it correctly predicts an incoherent abstract.
- False Positive (FP): When the model incorrectly predicts the positive class. This means it predicts an incoherent abstract as coherent.
- False Negative (FN): When the model incorrectly predicts the negative class. This means it predicts a coherent abstract as incoherent.

The basic components of the 6 experiments are detailed in the Table V:

TABLE V. EVALUATION COMPONENTS

Models	True Positive (TP)	False Positive (FP)	True Negative (TN)	False Negative (FN)
Experiment 1	118	70	79	32
Experiment 2	124	60	89	26
Experiment 3	114	55	94	36
Experiment 4	111	62	87	39
Experiment 5	137	83	66	13
Experiment 6	127	78	71	23

The Accuracy, F1-score, precision and recall were the most frequently used metrics to report model performance on benchmark datasets. As metrics for binary classification problems, they can be derived from a confusion matrix, a two by two contingency table of the predicted and observed class labels [40]. Once the evaluation components have been defined, performance measurements were calculated using the following equations: 7 8 9 y 10:

$$Precision = \frac{TP}{TP + FP} \quad (7)$$

$$Recall = \frac{TP}{TP + FN} \quad (8)$$

$$F1Score = 2 * \frac{Precision * Recall}{Precision + Recall} \quad (9)$$

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \quad (10)$$

Applying performance measurements equations to the 6 generated models, the following results were obtained from the Table VI.

TABLE VI. MODELS PERFORMANCE MEASURES

Models	Accuracy	Precision	Recall	F1-Score	Loss
Experiment 1	0.65	0.63	0.79	0.70	0.66
Experiment 2	0.71	0.67	0.83	0.74	0.55
Experiment 3	0.70	0.67	0.76	0.71	0.59
Experiment 4	0.66	0.64	0.74	0.69	0.62
Experiment 5	0.68	0.62	0.91	0.74	0.61
Experiment 6	0.68	0.62	0.84	0.71	0.64

The experiment 2 model offers a better Accuracy (0.71) to detect inconsistent sentences in scientific abstracts. This shows that performing a Second Fine-Tuned mixing data from the same language, but different domain (MESINESP + Comunicar), improves the Accuracy to evaluate coherence.

Experiment 3 shows that performing a Second Fine-Tuned by mixing data from a different language (English) and different domain (arXiv + ICMLA + Comunicar) improves accuracy to detect inconsistent sentences (0.70). This aspect is important, as it is shown that you can increase training data from different languages without worrying about results to evaluate coherence. This is because a multilingual pre-trained model of BERT has been used.

In experiments 4, 5 and 6 it is shown that the KRM method works slightly better than the RM method with few data. The KRM method, at first, better classifies incoherent abstracts,

but if a large set of data is increased whose negative examples were generated with RM, it does not offer higher performance. This happens because the testing data were not created with the KRM method but with the RM method, which negatively impacts predictions.

The authors agree with the work of Bendeviski *et al.* [30] when he mentions that BERT offers better results for the evaluation of coherence than traditional machine learning methods, since the latter cannot understand the context of a text as does BERT, in addition to that BERT offers greater scalability and guarantees the Transfer learning process.

In these studies [4], [6] and [23] they have focused on generating incoherent examples varying the order of sentences (Sentence Ordering Task) unlike the present research that has focused on the detection of inconsistent sentences for the evaluation of coherence in scientific abstracts written in English/Spanish.

V. CONCLUSION AND FUTURE WORK

In this study, it has been shown that abstracts written in different languages/domains can be trained to detect inconsistent sentences of test data whose language and domain is also different from training data. Experiment 2 has proved to be better for the detection of inconsistent sentences of abstracts written in Spanish. Experiment 3 has proved to be more optimal for the evaluation of abstracts written in Spanish using combined training and validation data written in Spanish and English. Also, the variety of incoherent abstracts generated with RM and KRM during the Second Fine-Tuned has proven to deliver no better results than to train with a single method of generating incoherent abstracts when you want to detect inconsistent sentences for coherence evaluation.

Future research will renew the current clustering method (KRM) to a BERT Score-based method for the detection of inconsistent sentences. It will also address the evaluation of coherence as a multi-classification problem taking into account the types of incoherence: contradiction, redundancy and thematic discontinuity.

ACKNOWLEDGMENT

To the Universidad Nacional de San Agustín de Arequipa for the funding granted to the project "Transmedia, Gamification and Video games to promote scientific writing in Engineering students", under Contract No. IBA-IB-38-2020-UNSA. We would like to thank to the "Research Center, Transfer of Technologies and Software Development R+D+i" – CiTeSoft-EC-0003-2017-UNSA, for their collaboration in the use of their equipment and facilities, for the development of this research work.

REFERENCES

- [1] Pishdad L., Fancellu F., Zhang R., Fazly A. "How coherent are neural models of coherence?", In Proceedings of the 28th International Conference on Computational Linguistics, Barcelona, Spain, pp. 6126–6138, December, 2020.
- [2] Charolles M, "Introduction aux problèmes de la cohérence des textes: Approche théorique et étude des pratiques pédagogiques", Langue française, 1978.

- [3] Xiong H., He Z., Wu H., Wang H., "Modeling coherence for discourse neural machine translation", In Proceedings of the AAAI Conference on Artificial Intelligence, Honolulu, HI, USA, Vol. 33, pp. 7338–7345, July, 2019.
- [4] Telenyk S., Pogorilyy S., Kramov A., "Evaluation of the Coherence of Polish Texts Using Neural Network Models", *Appl. Sci.*, April, 2021.
- [5] Andrade C., "How to write a good abstract for a scientific paper or conference presentation", *Indian J Psychiatry*, April, 2011.
- [6] Baiyun Cui, Yingming Li, Yaqing Zhang, and Zhongfei Zhang, "Text Coherence Analysis Based on Deep Neural Network", Proceedings of the 2017 ACM Conference on Information and Knowledge Management. 10.1145/3132847.3133047. October, 2017.
- [7] Peter W. Foltz, Walter Kintsch and Thomas K Landauer, "The measurement of textual coherence with latent semantic analysis", *Discourse Processes*, Vol. 25, 1998.
- [8] Md. Anwar Hussen Wadud and Md. Rashadul Hasan Rakib, "Text Coherence Analysis based on Misspelling Oblivious Word Embeddings and Deep Neural Network", *International Journal of Advanced Computer Science and Applications(IJACSA)*, 2021.
- [9] William C. Mann and Sandra A. Thompson, "Rhetorical structure theory: Toward a functional theory of text organization", *Text - Interdisciplinary Journal for the Study of Discourse*, Vol. 8, 1988.
- [10] Regina Barzilay and Mirella Lapata, "Modeling local coherence: An entity-based approach", *Association for Computational Linguistics*, Vol. 34, 2008.
- [11] Barbara J. Grosz, Scott Weinstein and Aravind K Joshi, "Centering: A framework for modeling the local coherence of discourse", *Association for Computational Linguistics*, Vol. 21, 1995.
- [12] Guinaudeau C., Strube M., "Graph-based Local Coherence Modeling", In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, Sofia, Bulgaria, Vol. 1, pp. 93-103, August, 2013.
- [13] Li J., Hovy E., "A Model of Coherence Based on Distributed Sentence Representation", In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, pp. 25–29, October, 2014.
- [14] J. Li and D. Jurafsky, "Neural net models for open-domain discourse coherence," *arXiv preprint arXiv:1606.01545*, 2016.
- [15] Putra J. W. G. and Tokunaga, T., "Evaluating text coherence based on semantic similarity graph", In Proceedings of TextGraphs-11: the Workshop on Graph-based Methods for Natural Language Processing, pp. 76–85, August, 2017.
- [16] Parikh A. P., Täckström O., Das D. and Uszkoreit J., "A decomposable attention model for natural language inference", In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, Austin, Texas, Association for Computational Linguistics, pp. 2249–2255, November, 2016.
- [17] Connor Holmes, Daniel Mawhirter, Yuxiong He, Feng Yan, Bo Wu, "GRNN: Low-Latency and Scalable RNN Inference on GPUs", In Fourteenth Eurosys Conference (Eurosys '19), March, 2019.
- [18] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakon Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin, "Attention is all you need", 31st Conference on Neural Information Processing Systems (NIPS), Long Beach, CA, USA, December, 2017.
- [19] Jacob Devlin and Ming-Wei Chnag, Research Scientists Google AI Language, "https://ai.googleblog.com/2018/11/open-sourcing-bertstate-of-art-pre.html", November, 2018.
- [20] Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova, "BERT:Pre-training of Deep Bidirectional Transformers for Language Understanding" Google AI Language, *arXiv:1810.04805v2 [cs.CL]*, 24 May, 2019.
- [21] Mohsen Mesgar, Michael Strube, "A Neural Local Coherence Model for Text Quality Assessment", In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, Association for Computational Linguistics, pp. 4328–4339, October, 2018.
- [22] Moon H. C., Mohiuddin T., Joty S. and Chi X., "A Unified Neural Coherence Model", Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, *arXiv e-prints*, 2019.
- [23] M. Bao, J. Li, J. Zhang, H. Peng and X. Liu, "Learning Semantic Coherence for Machine Generated Spam Text Detection", The 2019 International Joint Conference on Neural Networks (IJCNN), pp. 1-8 2019.
- [24] Elham Mohammadi, Timothe Beiko and Leila Kosseim, "On the Creation of a Corpus for Coherence Evaluation of Discursive Units", In Proceedings of the 12th Language Resources and Evaluation Conference, Marseille, France, European Language Resources Association, pp. 1067–1072, 2020.
- [25] Panitan Muangkammuen, Sheng Xu, Fumiyo Fukumoto, Kanda Runapongsa Saikaew, and Jiye Li, "A Neural Local Coherence Analysis Model for Clarity Text Scoring", In Proceedings of the 28th International Conference on Computational Linguistics, Barcelona, Spain (Online), International Committee on Computational Linguistics, pp. 2138–2143, 2020.
- [26] Callan D. and Foster, J., "Evaluation of Interest and Coherence in Machine Generated Stories", In *CEUR Workshop Proceedings*, Vol. 3105, pp. 212–223, 2021.
- [27] Q. H. Nguyen and M. Zaslavskiy, "Incoherent Sentence Detection in Scientific Articles in Russian and English," 2021 29th Conference of Open Innovations Association (FRUCT). 10.23919/FRUCT52173.2021.9435478. pp. 267-273, 2021.
- [28] D. J. Pierson, "The Top 10 Reasons Why Manuscripts Are Not Accepted for Publication", *Respiratory care*, October, 2004.
- [29] E.I. Bles and M.M. Zaslavskiy, "Criteria for text conformity to scientific style", *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, vol.19, pp.299-305, April, 2019.
- [30] Filip Bendeviski, Jumana Ibrahim, Tina Krulec, Theodore Waters, Nizar Habash, Hanan Salam, Himadri Mukherjee, and Christin Camia, "Towards Automatic Narrative Coherence Prediction", In Proceedings of the 2021 International Conference on Multimodal Interaction (ICMI '21), <https://doi.org/10.1145/3462244.3479895>, pp. 18-22, October, 2021.
- [31] Noji H. and Takamura H., "An Analysis of the Utility of Explicit Negative Examples to Improve the Syntactic Abilities of Neural Language Models", *arXiv e-prints*, 2020.
- [32] Comunicar, *Revista Científica de Comunicación y Educación*, <https://doi.org/10.3916/comunicar>, 2021.
- [33] Rana Ankush, Gonzalez-Agirre Aitor, Miranda-Escalada Antonio, and Krallinger Martin, "MESINESP: Medical Semantic Indexing in Spanish - Train dataset (1.0)", Zenodo, <https://doi.org/10.5281/zenodo.3826492>, 2020.
- [34] Sayak, "arXiv Paper Abstracts: arXiv paper abstract dataset for building multi-label text classifiers", Kaggle, <https://www.kaggle.com/datasets/spsayakpaul/arxiv-paper-abstracts>, 2021.
- [35] Vallejo Diego, Morillo Paulina, Ferri Cèsar, "ICMLA 2014/2015/2016/2017 Accepted Papers Data Set", Mendeley Data, V2, 10.17632/wj5vb6h9jy.2, 2019.
- [36] Souza F.D. and Filho J.B. de O. e S., "BERT for Sentiment Analysis: Pre-trained and Fine-Tuned Alternatives", *Computational Processing of the Portuguese Language*. https://doi.org/10.1007/978-3-030-98305-5_20. pp. 209–218, 2022.
- [37] Devlin J., Chang M. W., Lee K. and Toutanova, K., "Bert: Pre-training of deep bidirectional transformers for language understanding", In *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, Association for Computational Linguistics, Vol. 1, pp. 4171–4186, 2019.
- [38] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever and Ruslan Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting", *Journal of Machine Learning Research*.10.5555/2627435.2670313, January, 2014.
- [39] Bisong, E., "Building Machine Learning and Deep Learning Models on Google Cloud Platform", *Apress*. https://doi.org/10.1007/978-1-4842-4470-8_7, 2019.
- [40] Blagec K., Dorffner G., Moradi M. and Samwald M., "A critical analysis of metrics used for measuring progress in artificial intelligence", *arXiv preprint arXiv:2008.02577*, 2020.

Modeling and Simulation of Adaptive Traffic Control System for Multi-Intersection Management using Cellular Automaton and Queuing System

Salma EL BAKKAL, Abdallah LAKHOULI, El Hassan ESSOUFI
Hassan First University of Settat
Faculty of Sciences and Techniques
Mathematics, Computer Science and
Engineering Sciences Laboratory
MISI, 26000, Settat, Morocco

Abstract—During last years, urban traffic has become one of the most studied research topics. This is mainly due to the enlargement of the cities and the growing number of vehicles traveling in this road network. One of the most sensitive problems is to verify if the intersections are congestion-free. Another related problem is the automatic reconfiguration of the network without building new roads to alleviate congestions. These problems require an accurate model to determine the steady state of the traffic. The present article proposes an adaptive traffic light system based on the BCMP network queuing and cellular automata. The aim of this work is to predict the best red and green time span by combining three important factors: The queue length, the evacuation time and the capacity of the destination roads. This approach can maximize the number of vehicles passing intersection and at the same time can minimize the average waiting time of vehicles as a result reducing the congestion and keep the fluency in intersections. To validate our results, we compared our model with a fixed model to explain the strengths of our proposed algorithm.

Keywords—Traffic light systems; cellular automaton; BCMP; queuing systems; traffic congestion; waiting time; adaptive systems

I. INTRODUCTION

Traffic congestion is nowadays regarded as one of the biggest problems related to mobility in every country, mainly in big cities. This phenomenon causes many problems for people like lost time, fuel consuming and huge waste of energy. The inability to reduce traffic and the imbalance between the infrastructure and traffic demand are the main causes of the congestion. Therefore, the need to build new roads, bridges and tunnels are required, but it's necessary to combine with new technology in order to balance between the traffic demand and the existing infrastructure.

Many researches have been done in recent year in order to propose new technology to manage urban traffic smartly. The researchers concluded that if the intersections are occupied by vehicles, then traffic congestion occurs. And to solve this problem traffic light control system has been proposed in order to control the vehicles in the city, town and village. The most used traffic control systems are static, i.e., the time periods are given in advance as static results by calculating the delay time of the traffic lights using current situation with the help of sensors and cameras. But, in reality, the density in an

intersection is dynamic due to the instability of the traffic conditions in different time like rush hours. Therefore, the fixed strategy cannot match the need of the actual traffic situation. Thus, the intelligent traffic light systems have become an essential need in order to regulate the traffic by adjusting the time span based on the number and speed of the vehicles existing at the intersection in real time. Those systems can maximize the number of vehicles passing the intersection and minimize the number of the blocked one as a result, can effectively minimize the average waiting time of vehicles.

The most famous and successful adaptive traffic light system in the world is SCOOT [1] in England and SCATS [2] in Australia. The main objective of SCOOT system is minimizing the sum of the average queues in the area [3] and SCATS tries to find the best phasing for current traffic situation using a fixed plan [4]. Furthermore, various methods have been applied in the optimization traffic light protocol such as Fuzzy logic control [5] and [6] that use the Fuzzy logic and image processing to vary the timing of the traffic lights controllers. In [7] authors presents different artificial neural network approaches for computer to predict the traffic network. And [8] the authors propose a new algorithm to manage traffic light at the intersection using Genetic Algorithm. Also, cellular automata models are the most of microscopic models have been developed in recent year due to the flexibility and simplicity of modeling. The Biham, Middleton and Levin [9] was the first CA model applied to the urban traffic light, it describes the different state of traffic and identifies the key factors affecting the phase transition [10]. A lot of extensions of this model have been developed wish means that the cellular automata is a good way to describe the urban traffic flow. For example in [11] authors used a cellular automaton to simulate road traffic in order to present a novel application of array DBMSs. Otherwise in [12] authors proposed a new cellular automata model under Kerner's framework that manage vehicles by considering the effect of forward-backward vehicles in the internet of vehicles. In addition, authors of [13] used cellular automata to generate a mixed traffic flow in order to analyze the behavior of manually driven and autonomous vehicles. Otherwise, to predict the traffic situation and model highway traffic researchers use queuing models in order to propose the best set of time based on the queue length of each road in the intersection as result optimizing the traffic flow and evaluating

the system performance [14]. Most of methods developed in the recent period mainly focused on unsignalized intersections [15]. In [16] authors propose a review about queuing models of unsignalized intersections and in [17] authors presented an approach for both signalized and unsignalized intersections. The author in [18] presents an algorithm that identifies levels of congestion in traffic problems. The author in [19] propose a model for an urban road network dedicated to traffic intersection. And [20] represented an urban road intersection with the BCMP queuing network. Authors programmed a simulator to emulate the vehicle behavior at the intersection and they present a comparison with the analytical model and the proposed approach.

Most of the systems mentioned above focused only to optimize the average waiting time of vehicles in queues and it don't take into consideration the evacuation time of the crossroad. This later is an important parameter because if vehicles stay more time in the crossroad automatically the average waiting time of vehicles will increase and congestion will occur. And to solve this problem we need to guarantee the balance between the number of vehicles passing the intersection and the capacity of the destination roads.

The aim of this paper is to present a new adaptive traffic light system in which traffic light time changes in real time based on the queue length and the capacity of roads. To implement this model, the traffic roads have been designed with cellular automaton in order to study the vehicle behavior between intersections. And use the BCMP queuing system to model our network and calculate the performance measure such as the arrival rate, the average queues length and density of roads. The main objectives of our system are maximizing the number of passing vehicles, minimize the average waiting time in the network and ensure that all vehicles can leave the intersection in the least possible time.

The remaining of this paper is organized as follows: Section 2 presents the necessary scientific background about the BCMP queuing systems and the cellular automaton models. Section 3 presents our proposed approach. In Section 4 we will discuss the experimental results and its analysis, then in Section 5 the conclusion will be drawn.

II. BACKGROUND

A. BCMP Queuing System

BCMP is a queuing system that consists of a set of queuing centers or stations. Each service center has a scheduling discipline. Each client in the network has a class wish may influence the routing probabilities and the service time at the stations. The classes are labeled by $1, \dots, R$ and it can be partitioned into chains. The client enters the network from the outside according to a Poisson process with a give probability, waits in the queue for the service and it gets the next available center or exits the network. The BCMP network contains M stations and R client classes. Each client has a class and for each class a routing probabilities must be specified in order to describe the classes behavior throughout the network. A class can be open which means that clients of this class enters from the outside and eventually leave the network. Or closed which describes client that can never leave the network. Note that in this work only the open class has been studied. There are

TABLE I. EXAMPLE OF ROUTING PROBABILITIES

sectors	I_1	S_1	O_1
0	R_{0,I_1}	R_{0,S_1}	R_{0,O_1}
I_1	0	R_{I_1,S_1}	R_{I_1,O_1}
S_1	R_{S_1,I_1}	0	R_{S_1,O_1}
O_1	0	0	0

three sorts of service sectors in an open network. Input sector I_i used by clients to enter the network, output sector (O_i) used by clients to exit the network, and internal sector S_i used to move inside the network. Therefore, the routing probability is a float number R_{S_i,S_j} which describe the probability that a client can move from sector S_i to S_j . Table I presents an example of a routing probabilities values for a network composed by one input sector, one internal sector and one output sector. Note that the sector 0 describes the outsides of the networks. In BCMP network, queuing stations can be classified as one of the following:

- Type 1: The queuing discipline is first in first out (FIFO) and distribution of service time is exponential and class independent.
- Type 2: The queuing discipline is processor sharing.
- Type 3: All the stations have infinite servers (IS) wish means clients never wait in queue.
- Type 4: The service discipline is last in first out (LIFO) with preemptive resume.

And the following variables are used to describe the open BCMP network parameters:

- R: The number of traffic classes in the network,
- K_{ir} : The number of vehicles of the r^{th} class at the i^{th} sector.
- μ_{ir} : The service rate of the i^{th} sector of the r^{th} class.
- $R_{0,js}$: The probability in an open network that a vehicle from outside the network enters to the j^{th} sector of the s^{th} class.
- $R_{ir,0}$: The probability in an open network that a vehicle of the r^{th} class leaves the network after having been served at the i^{th} station.
- $\lambda_{0,ir}$: The arrival rate from outside to the i^{th} sector of the r^{th} class.
- λ_{ir} : The arrival rate of vehicle of the r^{th} class at the i^{th} sector.

The performance measures are described as following [20].

- ρ_{ir} : The utilization of the i^{th} sector by r^{th} class clients.

$$\rho_{ir} = \lambda_r * \frac{e_{ir}}{\mu_{ir}} \quad (1)$$

- k_{ir} : The average number of clients of r^{th} class at the i^{th} sector.

$$k_{ir} = \frac{\rho_{ir}}{1 - \rho_i} \quad (2)$$

- T_{ir} : The average response time of r^{th} class clients at the i^{th} sector.

$$T_{ir} = \frac{k_{ir}}{\lambda_{ir}} \quad (3)$$

- W_{ir} : The average waiting time of r^{th} class client at the i^{th} sector.

$$W_{ir} = T_{ir} - \frac{1}{\mu_{ir}} \quad (4)$$

- Q_{ir} : The average queue length of class r clients at the i^{th} sector.

$$Q_{ir} = \lambda_{ir} * W_{ir} \quad (5)$$

Note that i and j indicating sector numbers, while r indicates traffic class

Most researchers have applied the BCMP network system to solve problems related to urban traffic. In [21] authors present a queuing theoretic framework based on BCMP for modeling autonomous mobility on demand systems within capacitated road networks. The author in [22] approved using an open BCMP queuing network the uniqueness of solution to obtain an optimal static routing. In [23] authors proposed a simulation of parking using a network of service center capacity queues. in

B. Cellular Automaton

Cellular automata were originally proposed by John Von Neumann [24]. It consists of a grid of cells. These cells are all equal in size and the lattice can be finite or infinite in number of cells. Also, its dimension can be 1 wish called a linear string of cells, 2 wishes describe a grid of cells or even higher dimension. Every cellular automaton should have three elements:

- Cell's States: It's an integer represents the state of each cell.
- Cell's neighborhoods: To determine the evolution of the cell it should define neighborhoods for each cell. In the simplest case, for example, in a two-dimensional CA the four west, east, south and north adjacent cells are neighborhoods and their state can affect the state of cell in future steps.
- Transition function: It describes the rule followed so that the CA model evolves in time. This rule developed based on the neighborhoods state and model characteristic.

In recent years, most of the microscopic models developed by using the language of cellular automata (CA) [25]. Cremer and Ludwig [26] are the first whose proposed the first model of cellular automaton applied to the road traffic in 1986 and the Biham, Middleton and Levine model [9] is the first classical model applied to the urban traffic. This later describes the principal factors affecting phase transitions and identifies the different states of urban traffic flow [27]. A lot of researchers have proposed many extensions of the BML model. Fukui in 1996 [28] considered the average velocity in the BML and introduced the individual high-speed of vehicles by using the running speed of the traffic flow. Nagatani in 1995 [29] focused to reduce the numbers of gridlocks in the BML and improve

the running status of the traffic by using the cloverleaf junction. Cuesta (1993) [30] and Nagatani (1995) [29] were the first to propose switch rules in the BML model. Ding et al., 2011 [31] have explored the mean field theory in the BML model and the authors of [32] were the first to propose a modified BML to predict the urban traffic jams in real time.

III. PROPOSAL APPROACH

In this work, the adopted network contains four connected intersections managed by a set of traffic lights poles. An intersection formed by two intersecting perpendicular streets; each street contains bidirectional traffic roads. For the sake of simplicity, the vehicles in this paper can't change their direction which means that if a vehicle enters the network from the north side automatically will leaves the network from the south side. For more details, see Fig. 1.

The traffic lights in a given axis have the green light simultaneously, but they switch to the red at different times because our algorithm calculates time of the red or green color individually for each direction and to simplify more our model it supposed that there was no yellow light. In order to describe our network, the following parameters will be used in this paper:

- P_i : Priority of the green light that equal 1 if in the road i has higher priority to leave the intersection.
- GT : Green light time.
- RT : Red light time.
- ET : Evacuation time of the intersection.
- RC : Maximal capacity of the road.
- FS : Free space in front of each road near the intersection.
- N : The number of vehicles existing in the road.
- S : Light state Boolean that equal 1 if the light is green and 0 otherwise.

A. Proposed Cellular Automaton

In order to describe the representation of the traffic between the intersections a one-dimensional cellular automaton has been implemented. This later represent the road as a line of cells, each car occupied only one cell and each cell can host only one vehicle. All cars move in the same direction and we supposed in this work that lane changing is not allowed. Their positions are updated synchronously, in successive iterations (discrete time steps). Note that the required time for a vehicle to move from cell to another is one second. The motion rules of our model are simply that a car moves if its destination cell is empty and all the vehicles behind can move by one cell. Figure 2 shows an illustration of our transition rules.

In this work the space between all intersections is equal which mean that the cell number N in each road in our network is fixed.

The road density is calculated by the following equation:

$$\rho = \frac{\sum_{i=0}^N C_i}{N} \quad (6)$$

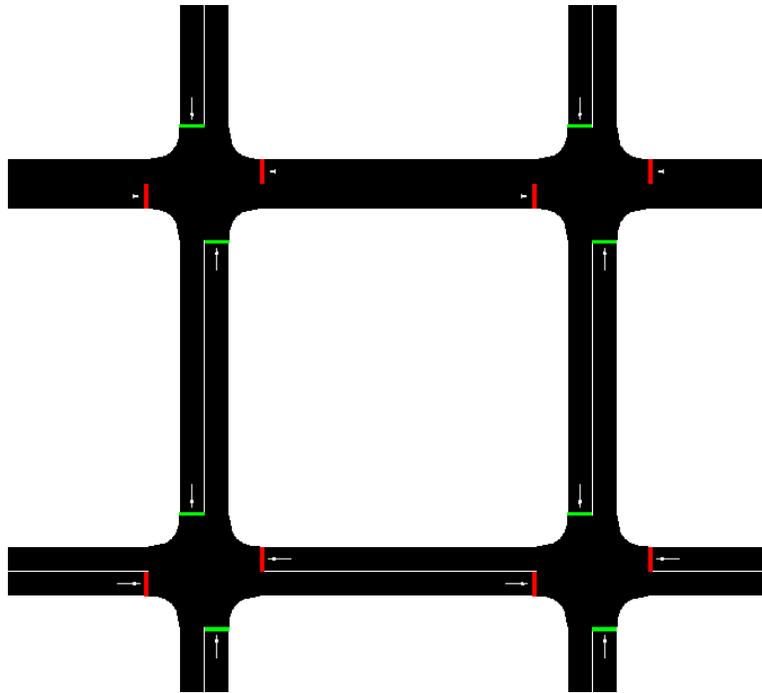


Fig. 1. Adopted Network.

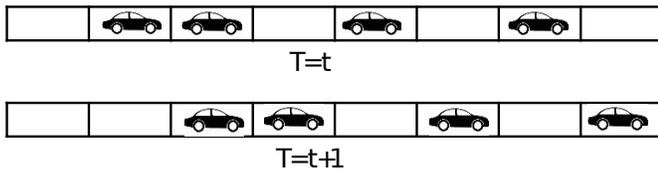


Fig. 2. Transition Rules.

Where C_i present a Boolean variable that equal 1 if the cell is occupied and 0 otherwise.

$$C_i = \begin{cases} 1 & \text{if occupied} \\ 0 & \text{else} \end{cases} \quad (7)$$

B. Proposed BCMP System

In this work the proposed open BCMP queuing network model composed with 8 inputs sectors (I_i), 8 outputs servers (O_i) and every queue in each intersection has internal sectors which means the proposed network have 16 internal sections ($S_{(i,j)}$). The arrival rates in this network following poisson distribution with parameter λ and the service rate following exponential distribution with parameter μ . The queues are type $M/M/1$ and the service policy is with first-come-first-served (FCFS).

The vehicle classes in this paper are defined as the flow of cars coming from the same input sector consequently 8 flow classes are identified and each class entering the network with parameter λ_i .

For example, class 1 is defined by the vehicles entering network through input I_1 with parameter λ_1 and leaving it

TABLE II. ROUTING PROBABILITIES OF THE VEHICLE CLASS 1

sectors	I_1	$S_{(1,1)}$	$S_{(4,1)}$	O_6
0	1	0	0	0
I_1	0	1	0	0
$S_{(1,1)}$	0	0	1	0
$S_{(4,1)}$	0	0	0	1
O_6	0	0	0	0

on sector O_6 . The internal sectors of this class are $S_{(1,1)}$ and $S_{(4,1)}$. See Fig. 3 for more details.

Our BCMP model have 8 traffic classes, for each class, a routing probabilities must be specified based on the previous rules. Vehicles are moved between any two stations according to given routing probabilities. For example (see table II) a class 1 traffic vehicle goes from the input to sector $S_{(1,1)}$ with probability $R_{I_1, S_{(1,1)}} = 1$ then it either move to sector $S_{(4,1)}$ with the probability $R_{S_{(1,1)}, S_{(4,1)}} = 1$ and it leaves the intersection from sector O_6 with probability $R_{S_{(4,1)}, O_6}$.

C. Proposed Algorithm

In this section the details of the adaptive control algorithm will be presented. The algorithm calculate the red/green time for each traffic signal using the following steps presented in algorithm 1.

The main objective of our model is to adapt the traffic signal duration with the intersection situation in order to minimize the average waiting of the vehicles in the network. The selection of the green time duration is based on the queue lane and the free space existing in front of the road (FS).

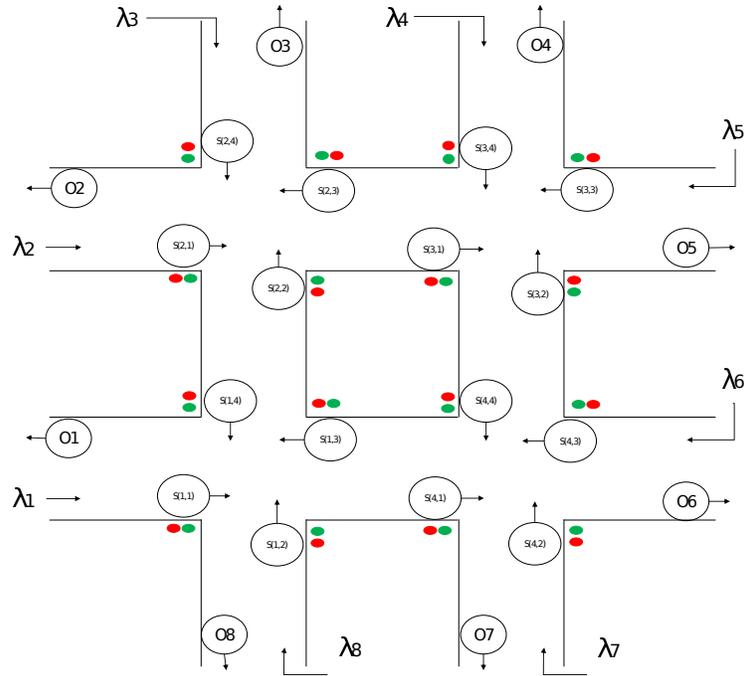


Fig. 3. Our BCMP Network.

Algorithm 1 Proposal Algorithm

- 1: **Procedure**
- 2: **FOR** $t=1 : T$ **do**
- 3: Take decision which axis has priority to have green time
- 4: calculate the green time using equation 10
- 5: generate the phase duration for all direction based on the decided green time
- 6: execute the phase
- 7: **END FOR**
- 8: **End procedure**

The selection process is performed every cycle and work as below: First from the intersection structure the axis that has priority to have the green time is known, and based on the waiting lane and free space, the algorithm calculates the green time for each direction with the equation below:

$$GT_{i,j} = \min(FS_{i,j}, Q_{i,j}) \quad (8)$$

and

$$FS_{i,j} = N - \sum_{k=0}^N C_{ik} \quad (9)$$

Where i describes the intersection number and j the direction. Note that in every intersection $j = 1$ and $j = 4$ define the horizontal axis and $j = 2$ and $j = 3$ define the vertical one. After the program calculate the green time for each direction, then, calculate the cycle duration by adding the evacuation time of the intersection (ET_i) to the max of the two green times selected. For example, if the horizontal

TABLE III. ARRIVAL RATE VALUE

λ values	Traffic state
0.2	free flow state
0.4	medium flow state
0.9	congested flow state

axis has the priority, the cycle duration is equal to the follow equation:

$$H_i = \max(GT_{i,1}, GT_{i,4}) + ET \quad (10)$$

IV. NUMERICAL SIMULATION

There are two simulated approaches, our model and the fixed model where the green and red time values are fixed. This simulation will lead to explain the differences between these two models and show the performance of the proposed adaptive system.

In this section the arrival and service rates parameters are presented.

A. Arrival Rates

This parameter defines the number of vehicles that arrive at a sector per time unit. As mentioned, we have 8 inputs which mean we have 8 arrival rate $[\lambda_1, \dots, \lambda_8]$.

In Table III we present the different arrival rate with the situation that they occurred.

B. Service Rates

This rate defines the number of vehicles crossing a sector per second, it supposed that all the sectors in our network have

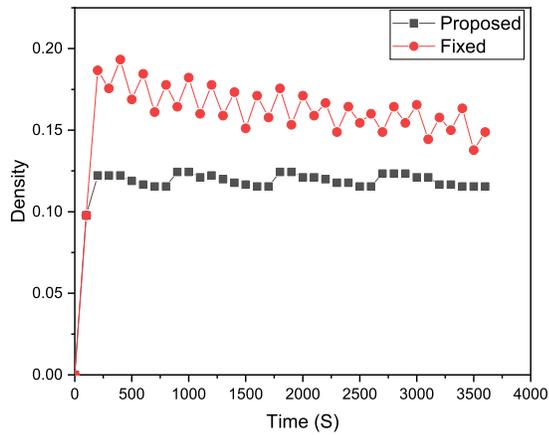


Fig. 4. Comparison of Density for Free Flow State.

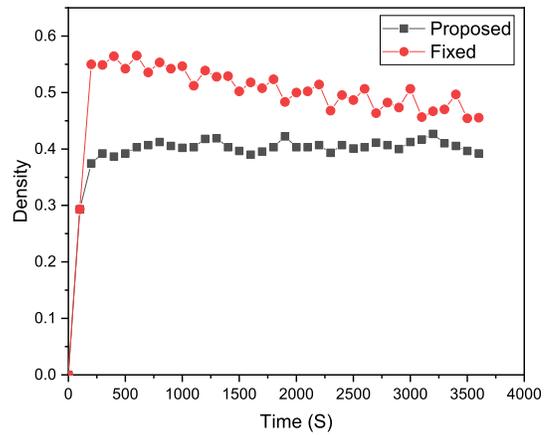


Fig. 6. Comparison of Density Congested Flow State Scenario.

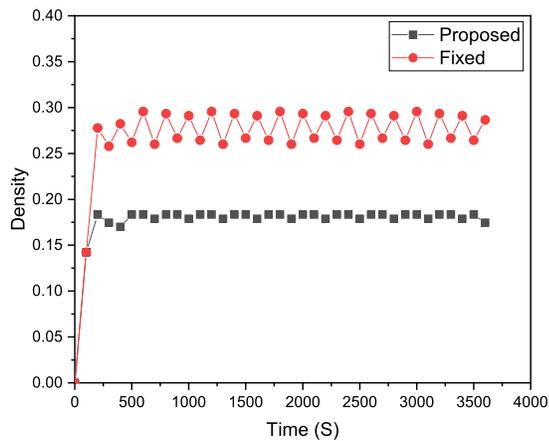


Fig. 5. Comparison of Density for Medium Flow State.

the same rate value $\mu = 1$. Note that in this paper, it supposed that the number of cells between the intersections is equal to 40 and the evacuation time is equal to 3 seconds.

V. RESULTS AND DISCUSSION

To study the performance of our proposed model, we need to observe the behavior of the network density during 1 hour with different arrival rate values. For free flow state, the density was on average of 0.12 using our proposed model and 0.16 with the fixed model. See Fig. 4.

For medium flow state, the density was on average of 0.18 using our proposed model and 0.27 the fixed model. See Fig. 5.

For the congested flow state, the density was on average of 0.4 using our proposed model and 0.5 with the fixed model. See figure 6.

The average values show which model guarantee more

fluent traffic and that impact directly the number of vehicles released from the network. From these figures, it can be observed that our proposed model allows vehicles to leave the network earlier than the fixed model Whereas after 1 hour our proposed model release 20% of vehicles with free flow state, 37% of vehicles with the medium flow state and 14% of vehicles with the congested flow state, more than the fixed model.

Fig. 7 presents a special simulation because in this scenario the arrival rate $\lambda = 1$ which mean every second one vehicle enter from every input of the network. and it can be clearly seen that after 500 sec the fixed system was blocked. This phenomenon in reality needs a police agent to regulate the traffic, which explain the importance of the free space parameter in the prediction of the cycle duration. And in order to show the difference between our model and the others model that based only on the queue length, we add a comparison between our model and the same configuration, but without the free space parameter and it's clearly presented in Fig. 7 that our model keeps the fluency of the traffic and reduce the probability to have jammed state.

These results demonstrate that our model is more efficient than the fixed strategy in term of the density. This is because our algorithm makes the decision to turn the green light on when there are vehicles staying in the intersection and turn off when the queue is empty. Otherwise the fixed algorithm turns the green light based on a static strategy whatever the queue length. Also in term of the traffic fluency, the average throughput under free, medium and congested traffic situation is shown in Fig. 8. This later presents that our model keeps the traffic smoothly by minimizing the difference between the arrival throughput and the output flow in different traffic states. Note that when this difference is minimized means that the arrival vehicles can easily enter the network and leave it with a minimal traveling time as shown in the next paragraphs.

In the next figures the difference between our algorithm and the fixed algorithm in term of the average queue length is presented. The average queue length of the first intersection was extracted in order to study the effect of our proposed

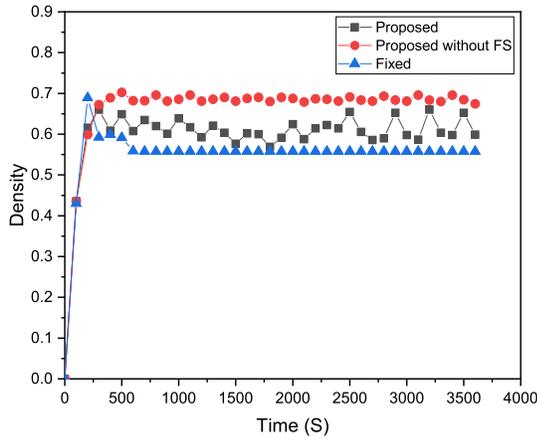


Fig. 7. Comparison of Density for Jammed Flow State.

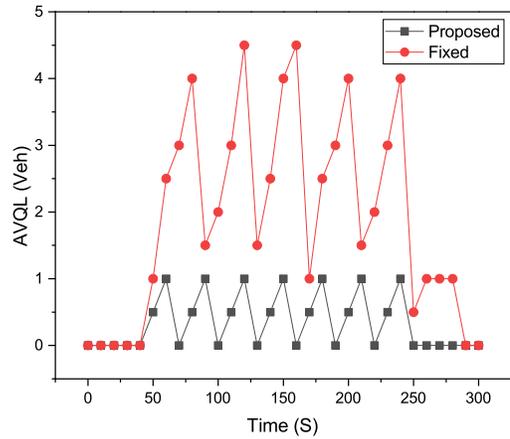


Fig. 9. Results of the Average Queue Length for Medium Flow State.

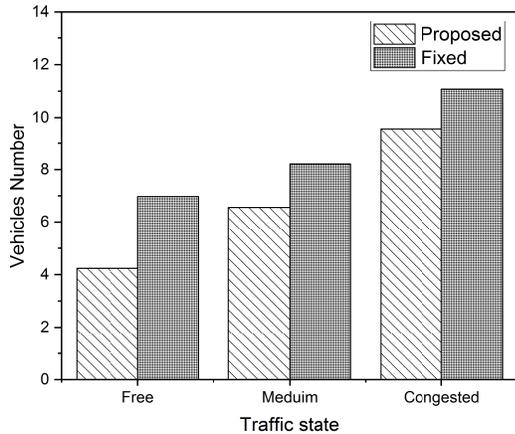


Fig. 8. Comparison of the Average Throughput.

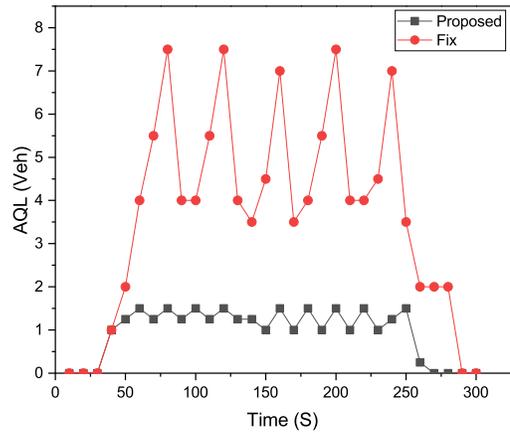


Fig. 10. Results of the Average Queue Length for Congested Flow State.

algorithm on different arrival rate.

In Fig. 9, the average queue length still between 0 and 1 with our proposal with an arrival rate $\lambda = 0.4$ otherwise with the fixed algorithm the average is between 1 and 4.5 which explain an important difference. Also, in congested flow state our model shows an important performance because the maximum of the average queue length was 1.5 where with the fixed strategy was 7.5. See Fig. 10. This can be explained by the fact that, when the queue length is considered to predict the green time, the average queue length decreases automatically.

From this, it can be seen that our algorithm guarantees to the vehicles in the intersection a low waiting time compared to the fixed algorithm. And in order to show more the performance of our system the average velocity in the network was studied in order to analyze the behaviors of the vehicles over time. Fig. 11 confirms that our model maximizes the average velocity in the network. But the most important fact deduced from this figure is that with the fixed algorithm the traffic

block earlier than with our model wish confirm the effect of minimizing the queue length in the intersections.

Fig. 12 presents the average waiting time in our network. The waiting time is defined as the time required for a vehicle to leave the network. In this simulation, it supposed that the arrival rates are equal in order to study the behavior of our model in different status. The results shows that the vehicles wait less with the proposed method under all traffic conditions. In the free traffic, the average waiting time values are 102 s and 180 s with proposed model and fixed model, respectively. Thus, with the medium and congested traffic, the vehicles wait less than 122 s and 187 s, respectively with our control method where with the fixed model wait 185 s and 268 s, respectively. And note that our model ensures the improvements 43%, 34%, and 31% in average waiting time values under free, medium and congested traffic, respectively.

Hence, our proposed model reduces the density, the average

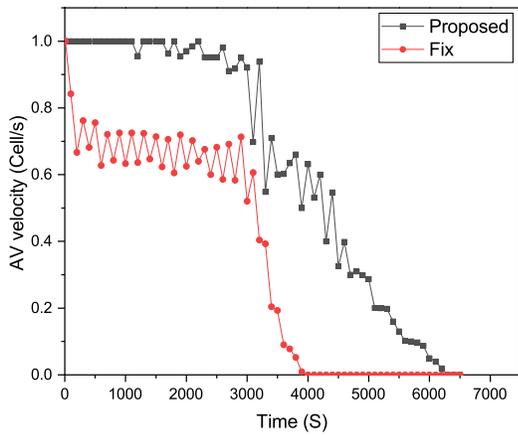


Fig. 11. Results of the Average Velocity.

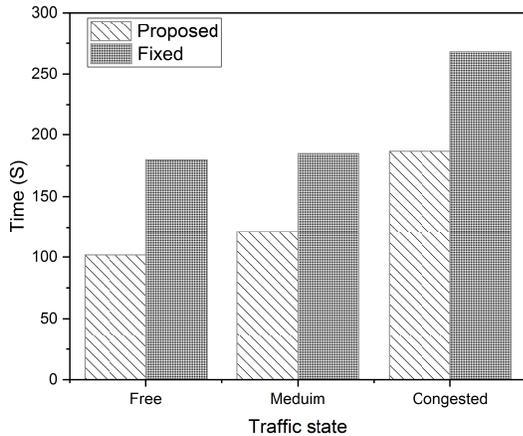


Fig. 12. Comparison of the Average Waiting Time.

queue length and the average and so it can maximize the average velocity of vehicles at the intersections.

After that and to improve more our model we consider in the next results that the arrival rate from the network inputs are not equal. We supposed that we have 4 type of arrival rates λ_{south} for the vehicles coming from the south of the network, λ_{north} for the vehicles coming from the north, λ_{east} for the vehicles coming from the east and λ_{west} for the vehicles coming from the west. And we implemented 4 scenarios. Firstly, the λ_{south} and λ_{north} are equal to 0.4 which implies that we have horizontally a medium flow and vertically we have a congested flow i.e. $\lambda_{west} = \lambda_{east} = 0.9$. This scenario aims to compare the difference between our model and the fixed model in term of the average queue length for every direction.

Note that, the following parameters are defined:

- AV : Average.

- NV : Number of Vehicles.
- $AVNQL$: Average Queue Length for North Vehicles.
- $AVSQL$: Average Queue Length for South Vehicles.
- $AVEQL$: Average Queue Length for East Vehicles.
- $AVWQL$: Average Queue Length for West Vehicles.

The Table IV shows a big difference between the two models, because our model turns off the green light when all the waiting vehicles in the south and north are served meaning that it give the priority to the others vehicle to be served earlier, while with the fixed model, the vehicles have to wait even if the other directions are empty. Which signifies the difference in term of the average velocity because with our model we can maximize the average velocity with 29% than the fixed model. The others scenarios supposed that vehicles come with a random arrival rate. As shown in the table, we notice that our model is better than the fixed model even if the situation. For example, in the fourth scenario our model maximizes the average velocity by 37% than the other model. Although sometimes there is not a big difference between the two models, for example, in the second scenario the difference between the average south queue length with our model and the fixed model is just 0.2. But our algorithm keeps the most interesting ones. And the reason theoretical of these results is due to the proposed new algorithm which is already explained above.

VI. DISCUSSION AND LIMITATION

Test results shows that the installed algorithm at the intersection ensures traffic flow in conditions of unsaturated flows. In addition, this demonstrates that our method can be successfully used for the construction of adaptive traffic control in conditions where traffic is related to time-varying traffic flows. In our optimization method, the traffic demands are well estimated based on the information from the adjacent intersections, which brings a good coordination among the intersections. It is worth to mention that our optimization method has a chance to obtain a better control scheme and to achieve an optimal solution for the network. However, the selection of the time based on the free space existing in the network affects the performance since using only the average queue length parameter might introduce longer unnecessary traffic delay to intersections just because the vehicles have not enough space in their destinations. Thus, in theoretical analyses, adapting the green time span using the combination between the queue length and the free space may reduce the delay time and keep the traffic fluency.

VII. CONCLUSION

In this paper, an intelligent system based on cellular automaton and BCMP queuing system has been proposed. The main advantage of this model is to avoid congestion at intersections due to the applied algorithm that calculate in every cycle time the free space on the roads and the waiting line in order to give the appropriate green time for every intersection taking into account the priority and intersection evacuation time. The results prove that our approach keeps the traffic fluency at the intersections and reduce the probability to attempt jammed situation than fixed method. Using BCMP

TABLE IV. THE OBTAINED RESULTS FOR DIFFERENT SCENARIOS

Scenario	Approach	λ_{north}	λ_{south}	λ_{east}	λ_{west}	AV Velocity	AV NV	AVNQL	AVSQL	AVEQL	AVWQL
Scenario 1	Proposed	0.4	0.4	0.9	0.9	0.97	106.76	1.79	0.65	2.95	2.85
Scenario 1	Fixed	0.4	0.4	0.9	0.9	0.75	125.33	3.13	3.13	19.43	19.29
Scenario 2	Proposed	0.4	0.9	0.4	0.9	0.76	130.46	2.77	19.15	2.76	19.32
Scenario 2	Fixed	0.4	0.9	0.4	0.9	0.65	146.2	6.59	19.38	5.86	20.37
Scenario 3	Proposed	0.2	0.9	0.2	0.9	0.76	119.18	1.33	3.85	19.38	19v53
Scenario 3	Fixed	0.2	0.9	0.2	0.9	0.62	141.38	14.7	17.1	19.66	20.37
Scenario 4	Proposed	0.4	0.2	0.4	0.2	0.98	43.73	0.54	0.59	0.29	0.467
Scenario 4	Fixed	0.4	0.2	0.4	0.2	0.71	58.04	2.77	2.76	1.62	1.6

performance measures helps us to manage the queues in our network, which guarantee an optimal average waiting time for vehicles.

Note that this approach can not only applied in transportation domain it can also apply in different situation like bank service, hospital out-patient service, etc.

In the future works, we will integrate other technologies like machine learning and multi-agent system to develop more our approach and guarantee a good vehicle experience in cities.

REFERENCES

[1] P. Hunt, D. Robertson, R. Bretherton, and R. Winton, "Scoot-a traffic responsive method of coordinating signals," *Transport and Road Research Lab., Crowthorne, UK*, 1981.

[2] P. Lowrie, "Scats, sydney co-ordinated adaptive traffic system: A traffic responsive method of controlling urban traffic," 1990.

[3] D. I. Robertson and R. D. Bretherton, "Optimizing networks of traffic signals in real time-the scoot method," *IEEE Transactions on vehicular technology*, vol. 40, no. 1, pp. 11–15, 1991.

[4] J. Li, Y. Zhang, and Y. Chen, "A self-adaptive traffic light control system based on speed of vehicles," in *2016 IEEE International Conference on Software Quality, Reliability and Security Companion (QRS-C)*. IEEE, 2016, pp. 382–388.

[5] A. H. Pohan, L. A. Latiff, R. A. Dziauddin, and N. H. A. Wahab, "Mitigating traffic congestion at road junction using fuzzy logic," *International Journal of Advanced Computer Science and Applications*, vol. 12, no. 8, pp. 293–299, 2021. [Online]. Available: www.scopus.com

[6] A. Chabchoub, A. Hamouda, S. Al-Ahmadi, and A. Cherif, "Intelligent traffic light controller using fuzzy logic and image processing," *International Journal of Advanced Computer Science and Applications*, vol. 12, no. 4, pp. 396–399, 2021, cited By :2. [Online]. Available: www.scopus.com

[7] T. P. Oliveira, J. S. Barbar, and A. S. Soares, "Computer network traffic prediction: a comparison between traditional and deep learning neural networks," *International Journal of Big Data Intelligence*, vol. 3, no. 1, pp. 28–37, 2016.

[8] A. Merbah and A. Makrizi, "Optimal management adaptive of two crossroads by genetic algorithm," *International Journal of Modeling, Simulation, and Scientific Computing*, vol. 10, no. 03, p. 1950009, 2019.

[9] O. Biham, A. A. Middleton, and D. Levine, "Self-organization and a dynamical transition in traffic-flow models," *Physical Review A*, vol. 46, no. 10, p. R6124, 1992.

[10] W. Hu, L. Yan, H. Wang, B. Du, and D. Tao, "Real-time traffic jams prediction inspired by biham, middleton and levine (bml) model," *Information Sciences*, vol. 381, pp. 209–228, 2017.

[11] R. A. Rodrigues Zalipynis, "Convergence of array dbms and cellular automata: A road traffic simulation case," in *Proceedings of the 2021 International Conference on Management of Data*, 2021, pp. 2399–2403.

[12] H.-T. Zhao, L. Lin, C.-P. Xu, Z.-X. Li, and X. Zhao, "Cellular automata model under kerner's framework of three-phase traffic theory considering the effect of forward-backward vehicles in internet of

vehicles," *Physica A: Statistical Mechanics and its Applications*, vol. 553, p. 124213, 2020.

[13] X. Hu, M. Huang, and J. Guo, "Feature analysis on mixed traffic flow of manually driven and autonomous vehicles based on cellular automata," *Mathematical Problems in Engineering*, vol. 2020, 2020.

[14] M. Papageorgiou, C. Diakaki, V. Dinopoulou, A. Kotsialos, and Y. Wang, "Review of road traffic control strategies," *Proceedings of the IEEE*, vol. 91, no. 12, pp. 2043–2067, 2003.

[15] C. Osorio and M. Bierlaire, "A surrogate model for traffic optimization of congested networks: an analytic queueing network approach," *Tech. Rep.*, 2009.

[16] D. Heidemann and H. Wegmann, "Queueing at unsignalized intersections," *Transportation Research Part B: Methodological*, vol. 31, no. 3, pp. 239–263, 1997.

[17] D. Heidemann, "A queueing theory approach to speed-flow-density relationships," in *TRANSPORTATION AND TRAFFIC THEORY. PROCEEDINGS OF THE 13TH INTERNATIONAL SYMPOSIUM ON TRANSPORTATION AND TRAFFIC THEORY, LYON, FRANCE, 24-26 JULY 1996*, 1996.

[18] A. Lozano, G. Manfredi, and L. Nieddu, "An algorithm for the recognition of levels of congestion in road traffic problems," *Mathematics and Computers in Simulation*, vol. 79, no. 6, pp. 1926–1934, 2009.

[19] A. D Ambrogio, G. Iazeolla, L. Pasini, and A. Pieroni, "Simulation model building of traffic intersections," *Simulation Modelling Practice and Theory*, vol. 17, no. 4, pp. 625–640, 2009.

[20] J. Dad, M. Ouali, and Y. Lebbah, "A multiclass bcnp queueing modeling and simulation-based road traffic flow analysis," *World Academy of Science, Engineering and Technology*, pp. 78–2011, 2011.

[21] R. Iglesias, F. Rossi, R. Zhang, and M. Pavone, "A bcnp network approach to modeling and controlling autonomous mobility-on-demand systems," *The International Journal of Robotics Research*, vol. 38, no. 2-3, pp. 357–374, 2019.

[22] H. Kameda and Y. Zhang, "Uniqueness of the solution for optimal static routing in open bcnp queueing networks," *Mathematical and Computer Modelling*, vol. 22, no. 10-12, pp. 119–130, 1995.

[23] C. Dowling, T. Fiez, L. Ratliff, and B. Zhang, "How much urban traffic is searching for parking? simulating curbside parking as a network of finite capacity queues," *arXiv preprint arXiv:1702.06156*, 2017.

[24] J. Neumann, A. W. Burks *et al.*, *Theory of self-reproducing automata*. University of Illinois press Urbana, 1966, vol. 1102024.

[25] P. Sarkar, "A brief history of cellular automata," *Acm computing surveys (csur)*, vol. 32, no. 1, pp. 80–107, 2000.

[26] M. Cremer and J. Ludwig, "A fast simulation model for traffic flow on the basis of boolean operations," *Mathematics and computers in simulation*, vol. 28, no. 4, pp. 297–303, 1986.

[27] E. Brockfeld, R. Barlovic, A. Schadschneider, and M. Schreckenberg, "Optimizing traffic lights in a cellular automaton model for city traffic," *Physical review E*, vol. 64, no. 5, p. 056132, 2001.

[28] M. Fukui and Y. Ishibashi, "Traffic flow in 1d cellular automaton model including cars moving with high speed," *Journal of the Physical Society of Japan*, vol. 65, no. 6, pp. 1868–1870, 1996.

[29] T. Nagatani, "Self-organization in 2d traffic flow model with jam-avoiding drive," *Journal of the Physical Society of Japan*, vol. 64, no. 4, pp. 1421–1430, 1995.

- [30] J. A. Cuesta, F. C. Martínez, J. M. Molera, and A. Sánchez, "Phase transitions in two-dimensional traffic-flow models," *Physical Review E*, vol. 48, no. 6, p. R4175, 1993.
- [31] Z.-J. Ding, R. Jiang, W. Huang, and B.-H. Wang, "Effect of randomization in the biham-middleton-levine traffic flow model," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2011, no. 06, p. P06017, 2011.
- [32] W. Hu, L. Yan, H. Wang, B. Du, and D. Tao, "Real-time traffic jams prediction inspired by biham, middleton and levine (bml) model," *Information Sciences*, vol. 381, pp. 209–228, 2017.

Non-Parametric Stochastic Autoencoder Model for Anomaly Detection

Raphael Alampay[✉], Patricia Angela Abu[✉]
Ateneo Laboratory for Intelligent Visual Environment
Dept. of Information Systems and Computer Science
Ateneo de Manila University
Quezon City, Philippines

Abstract—Anomaly detection is a widely studied field in computer science with applications ranging from intrusion detection, fraud detection, medical diagnosis and quality assurance in manufacturing. The underlying premise is that an anomaly is an observation that does not conform to what is considered to be normal. This study addresses two major problems in the field. First, anomalies are defined in a local context, that is, being able to give quantitative measures as to how anomalies are categorized within its own problem domain and cannot be generalized to other domains. Commonly, anomalies are measured according to statistical probabilities relative to the entire dataset with several assumptions such as type of distribution and volume. Second, the performance of a model is dependent on the problem itself. As a machine learning problem, each model has to have parameters optimized to achieve acceptable performance specifically thresholds that are either defined by domain experts or manually adjusted. This study attempts to address these problems by providing a contextual approach to measuring anomaly detection datasets themselves through a quantitative approach called *categorical measures* that provides constraints to the problem of anomaly detection and proposes a robust model based on autoencoder neural networks whose parameters are dynamically adjusted in order to avoid parameter tweaking on the inferencing stage. Empirically, the study has conducted a relatively exhaustive experiment against existing and state of the art anomaly detection models in a semi-supervised learning approach where the assumption is that only normal data is available to provide insight as to how well the model performs under certain quantifiable anomaly detection scenarios.

Keywords—Neural networks; autoencoders; machine learning; anomaly detection; semi-supervised learning

I. INTRODUCTION

Anomaly detection is a widely studied field in computer science dating back to 1887 (Edgeworth) with applications ranging from intrusion detection, fraud detection, medical diagnosis, quality assurance and manufacturing. Anomalies / outliers / novelties are observations that exhibit characteristics that are not part of the usual pattern or expected behavior of what are considered as normal observations. These often result from an erroneous recording of information or fault in producing the information or in some cases, an intended act of disruption as with the case of intrusion in a computerized network system. The definition of what is normal however and consequently what constitutes to an anomaly, is largely dependent on the context of the domain being observed or practiced. For example, medical diagnosis might yield a relatively larger magnitude of deviation to consider something to be malignant

rather than benign compared to quality assurance in manufacturing where a relatively smaller magnitude of deviation is observed to consider something to be acceptable or not. In this case, it is a simplistic definition of a large bias occurring or an extremely uneven ratio that defines anomalies. Often, this magnitude of deviation is defined by a domain expert in order to determine the impact of identifying anomalies. Otherwise, the measurement of deviation to define normal from anomalies is empirically defined through numerous experimentation to determine an acceptable value for discrimination which is also influenced by either policy or industry standards. From this, we can say that the study of outlier detection and its applications are generalized as a binary classification problem where observations are categorized as either normal or anomalous and that it is contextualized within the domain at hand. Its value in an operational or business perspective is that the identification of anomalies would always result in actionable items to either fix a system or process to improve the overall output of the application [1]. Mathematically, anomaly detection can be expressed in general using the following equation:

$$f(x) > t \quad (1)$$

where t is some threshold value, x is some unknown data point and $f(x)$ gives a score for the data point. If the score falls above (or below depending on context) of the threshold t then x is considered to be an outlier. The value of t is defined *a priori* usually by a domain expert or through experimentation.

II. RELATED WORKS

A. Anomaly Detection

The study of anomaly detection can be generalized as a binary classification problem with labels *normal* and *outliers*. As a classification problem, observations whether labelled normal or anomalous are characterized with a fixed set of attributes. Anomalies are observations whose attributes deviate from normal data based on some acceptable magnitude. In general, anomalies comprise of an extreme minority of the overall data having very low occurrences. Other terms for anomalies are *outliers*, *novelties* or *abnormalities*. Regardless of normal or anomalous data, these observations can either be univariate or multivariate in nature.

B. Types of Anomalies

Depending on the problem at hand, an anomaly can be described in either of three major categories – point, contextual or collective anomalies.

1) *Point Anomalies*: Point anomalies are the simplest type of anomalies that are described as a single data point in n-dimensional space (regardless of it being univariate or multivariate data). Each data point weather anomalous or normal exist based on the values of its attributes. Point anomalies are also the easiest to visualize as they are simply points in the search space of the problem. For example, Fig. 1 illustrates data as point anomalies in the waveform dataset:

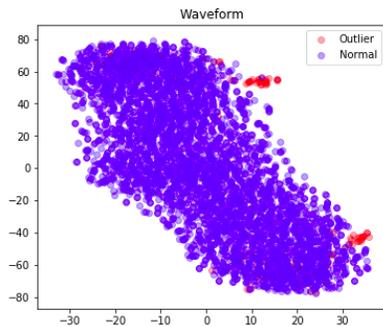


Fig. 1. Point Anomalies Example (TSNE Waveform)

Red points are considered anomalies if they deviate from majority of the normal data points (blue). Normal point data often cluster together as they exist more and exhibit similar measurements compared to anomalous observations. Thus this cluster of normal observations tend to form a cluster boundary. Points that lie beyond the boundary are considered to be outliers.

2) *Contextual Anomalies*: Also known as conditional anomalies, contextual anomalies are that which are bound to a specific context. A simple example of this would be the context of time. An outlier can be defined based on its measurement occurring at a specific point in time. When formulating an anomaly detection problem, it is integral to first define if the data can be simply described by its attributes (thus it is can be categorized as point anomalies) or if it can be contextualized by some dependent variable making it a contextual anomaly.

3) *Collective Anomalies*: Collective anomalies are defined to be a grouping of related data instances in relation to the entire data set. Individual observations may not be considered anomalies but if grouped together to form a higher level of observation, then we say that it is a collective anomaly. Defining collective anomalies are approached differently depending on the problem at hand. For example, in intrusion detection, a single observed usage of a protocol in a network may not be considered anomalous but if done in a certain sequence (a collection of protocol usage), such as network packets that use http, then ssh, then ftp protocols can be considered an attack vector (anomalies in this sense are defined as intrusions or attacks in the network).

C. Measuring Anomalies

As of this writing, there is no concrete definition of anomalies or anomaly detection datasets that distinguishes itself from any other binary classification problem. Furthermore, defining what anomalies are in a dataset are subject to the concrete problem at hand. Emott et. al however has proposed a set of measures to define how anomalies are measured in a local context. In their paper *Systematic Construction of Anomaly Detection Benchmarks from Real Data*, their study proposed four quantitative measures for defining anomalies relative to nominal data [2]. Of the four, three were implemented. Given a "parent" or "mother" dataset, it is possible to derive anomaly detection subsets based on difficulty constrained with a K parameter relating to the intended ratio of anomalies against a dataset. The three implemented measures for a given outlier data point are as follows. The fourth measure, *Feature Relevance* wasn't included in the paper and was open for interpretation and implementation. A portion of this study contributes to the implementation of this measure and the reasoning for the chosen approach. In depth discussion on the computation for the implemented scores are discussed in the methodology section.

D. Methods in Anomaly Detection

Defining a solution or training a model for anomaly detection can be categorized as either parametric or non-parametric. Regardless of approach, each category has its own share of advantages and disadvantages. Both categories are considered to be statistical approaches in anomaly detection.

1) *Parametric Models*: Anomaly detection models are considered to be parametric models if the problem assumes that the data being observed or generated follow a specific distribution. There are two key components for parametric model approaches. First would be the parameters θ for the assumed distribution. Second would be the probability density function $f(x, \theta)$, where x is an observation. The goal of parametric models is to derive or estimate the values of the parameters from the data itself where the parameters are largely dependent on the type of distribution assumed. The probability density function then outputs a score depicting how fit some observation x is given the estimated parameters of the distribution.

An example of a parametric model would be Gaussian based models. In this example, the structure of the data assumes to be gaussian in nature (exhibiting the normal distribution) that is, a symmetrical distribution where the mean is central and that data nearer to the mean occurs more frequently than data farther away from the mean exhibiting a bell shaped curve.

The parameters θ of the gaussian distribution are $\{\mu, \sigma^2\}$ which can be derived from training data x using equations 3, 4, 5 and 6. Given some unknown observation, we can check its probability score by fitting its attributes to the distribution's probability density function. The anomaly detection model can now be expressed using equation 2 wherein the unknown data \hat{x} is considered anomalous if its probability value falls below some defined threshold t .

$$f(\hat{x}|\theta) < t \quad (2)$$

$$\forall x \in \{1 \dots I\} x_{\mu_i} = \mu_i = \frac{\sum_{n=1}^N x_i}{N} \quad (3)$$

$$\forall x \in \{1 \dots I\} x_{\sigma_i} = \sigma_i^2 = \frac{\sum_{n=1}^N (x_i - x_{\mu_i})^2}{N} \quad (4)$$

$$f(\hat{x}_i | \mu_i, \sigma_i^2) = \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left(-\frac{(\hat{x}_i - \mu_i)^2}{2\sigma_i^2}\right) \quad (5)$$

$$f(\hat{x}, \mu, \sigma^2) = \forall \hat{x} \in \{1 \dots I\} \prod_{i=1}^I f(\hat{x}_i | \mu_i, \sigma_i^2) \quad (6)$$

As opposed to parametric models, non-parametric models allow the data to determine the underlying structure and boundaries for classification. No distribution is defined *a priori* unlike parametric models where it is largely based on an assumed distribution and its respective parameters.

2) *Classical Outlier Detection Models*: The following models are known to be used for outlier detection in literature and have served as standard benchmarks for newer models:

- Isolation Forest [3]
- Local Outlier Factor [4]
- Robust Covariance [5]

These models will be considered as part of performance measures against this study's proposed model.

3) *One-Class Support Vector Machines*: The One-Class SVM classification method is an extension of the original support vector machine classification algorithm as developed by Vapnik [6]. This approach however does not require data to be labeled as the algorithm returns a function that samples a small region from the probability distribution of the data that serves as the probability density of the training data. Because of this, One-Class SVM is categorized under the parametric class of anomaly detection algorithms. The function returns +1 for data points within the subregion and -1 elsewhere. The minimization function of One-Class SVM is slightly different from the original function and is characterized by equation 7.

$$\min_{w, \xi_i, \rho} \frac{1}{2} \|w\|^2 + \frac{1}{Cn} \sum_{i=1}^n \xi_i - \rho \quad (7)$$

subject to:

$$(w \cdot \phi(x_i)) \geq \rho - \xi_i \text{ for all } i = 1, \dots, n$$

$$\xi_i \geq 0 \text{ for all } i = 1, \dots, n$$

where ϕ is the kernel function to project data to a higher dimensional space, ξ_i are slack variables and C , as opposed to the original, decides the smoothness giving it a solution that a) sets an upper bound on the fraction of outliers and b) sets a lower bound on the number of training examples considered as support vectors. The solution creates a hyperplane characterized by w and ρ which has the maximum distance from the point of origin of the feature space and separates the data

points from the origin. As such, this can be categorized as an unsupervised learning algorithm since it takes into account all data points regardless of label. The mathematical equations for this algorithm are explained more in detail in [7].

4) *Autoencoders for Anomaly Detection*: Autoencoders are neural network models whose output is the same as its input. It approximates how to reconstruct the input by first compressing the data into lower dimensional space to represent a more generalized version of the input in what is called the encoding process. The lower dimensional representation of data, otherwise known as the latent layer as seen in Fig. 2, is then forwarded to the output layer which has the same dimensionality as the input in what is called the decoding process. thus this neural network looks for correlations of features in a data set while taking advantage of non-linear properties of neural networks.

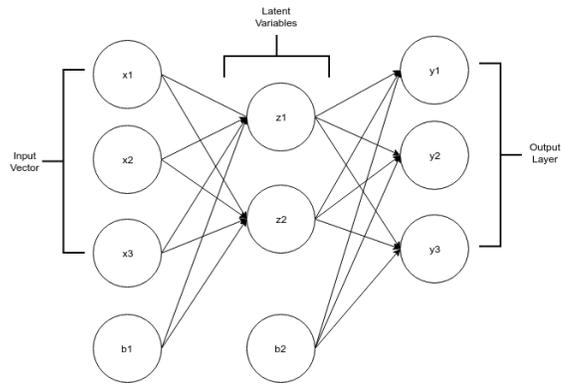


Fig. 2. Autoencoder Neural Network

Mathematically, the approximation of some input data x by the autoencoder model can be expressed in equation 8.

$$f(x) \approx x \quad (8)$$

During training, the error score of some input is the distance between the original input and resulting output. This is otherwise referred to as the reconstruction error. The distance or loss function commonly used is the mean of squared errors as computed in equations 9 and 10.

$$\text{MSE}(x, y) = \text{MSE}(x, f(x)) \quad (9)$$

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^n (x_i - y_i)^2 \quad (10)$$

The MSE value expresses the difficulty in reconstruction where the data fed to the network does not conform to what the model has learned from normal data. Thus if the value exceeds some threshold t , it can be concluded to be an outlier. This can be expressed in equation 11.

$$\text{MSE}(x, f(x)) > t \quad (11)$$

Applications of autoencoders in anomaly detection has been used extensively in the field of performance assessment in computing systems such as [8]. It has also been applied to numerous outlier benchmarking datasets where outlier ratio for validation falls below 30%. This can be seen in [9] wherein Yoshiao et. al proposed a method for estimating reconstruction capabilities of the autoencoder by disregarding high reconstruction errors produced by the model during mini-batch training resulting in partially selecting sets of training results to update the model during back propagation. This however only attempts to address the efficiency of training the autoencoder models while still maintaining a high enough accuracy as compared to standard autoencoders, autoencoder ensembles as well as One-Class SVM.

5) *One-Class Neural Networks*: The One-Class Neural Network as popularized by Chalapathy et. al., is typical feed forward neural network with a single node as its output [10]. However, this method's novelty can be described in its objective function which takes inspiration from One-Class SVM as well as its utilization of autoencoders. The objective function can be described in equation 12.

$$\min_{w, V, r} \frac{1}{2} \|w\|_2^2 + \frac{1}{2} \|V\|_F^2 + \frac{1}{v} \cdot \frac{1}{N} \sum_{n=1}^N \max(0, r - (w, g(VX_n))) - r \quad (12)$$

The key insight of the objective function is to replace the dot product from One-Class SVM's $(w, \phi(X_n))$ with the dot product $(w, g(VX_n))$ where V is the weight matrix from input to hidden layer that is optimized using an autoencoder. The change here allows transfer learning from an autoencoder model to be able to learn how the data points are reconstructed before applying it to a feed forward neural network. The values derived from the autencoder w and V are then used to optimize r which is theoretically the v -quantile of the array $(w, g(Vx_n))$. After getting the value of r , a score can be derived using equation 13:

$$S_n = \hat{y}_n - r \quad (13)$$

where \hat{y}_n is the output of the feed forward network for data point n . If S_n is greater than or equal to 0 then x_n is considered normal. Else, the point is said to be anomalous.

E. Well Known Autoencoder based Anomaly Detection Models

1) *Unsupervised Novelty Detection using Deep Autoencoders with Density based Clustering*: Deep autoencoders with density based clustering (DADBC) is a system developed by Amarbayasgalan et. al. to solve the anomaly detection problem in a fully unsupervised fashion. The main steps of the method is a) dimensionality reduction and b) novelty identification through clustering [11]. An illustration of the system can be seen in Fig. 3.

F. Problems in Anomaly Detection

In this research, we try to address key problems in anomaly detection. These have been identified both in literature as well as observations in the nature of existing solutions weather it

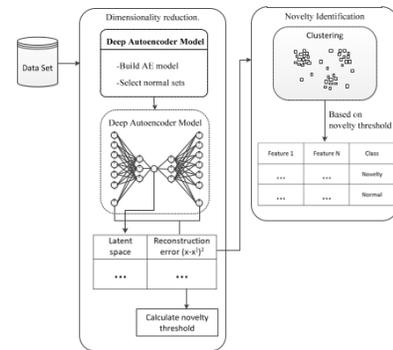


Fig. 3. Deep Autoencoders with Density based Clustering

be parametric or non-parametric. The fundamental problems are identified as follows:

- 1) If the model for anomaly detection contains parametric properties, there is heavy reliance on the assumed statistical distribution the data takes up.
- 2) Anomaly detection applications are largely contextual. In a non-parametric statistical setup, configuration of bins and clusters to profile data points have to be defined by a domain expert or empirically validated through experimentation. A wide variety of anomaly detection models require the definition of some threshold parameter to indicate the magnitude of variety between normal and anomalous data. This threshold is often defined by the practitioner or expert.
- 3) Anomaly detection datasets have a significantly low volume of data that are considered to be anomalous. In the preliminary results of this study, available standard anomaly detection datasets range from as little as 0.4% to at most 30% of data as anomalies making supervised learning techniques not viable. As such, this research only considers semi-supervised and unsupervised models for anomaly detection.

III. METHODOLOGY

This study's methodology is divided into two main parts. The first part contains the processes involved in defining quantitative measures of anomaly detection datasets which is largely based on Emott et. al's work on systematic construction of anomaly benchmarks. The main differences are the implementation of the feature relevance metric, removal of the constraint on selecting a user defined K relative frequency value and providing a global semantic variation score considering multidimensional nature of the data points. This modified approach in providing quantifiable features for anomaly detection datasets benchmarks is referred to as *categorical measures*. The second part of the methodology involves the construction of a neural network autoencoder based approach in solving the problem of anomaly detection. The novelties of the model lie in its non-parametric approach in defining a threshold value during inferring allowing it to adapt to the patterns exhibited by the data set itself and improving its performance with a stochastic component and a new loss function that prevents it from overfitting.

A. Categorical Measures

Given a data set, Emott et. al defined a set concrete measurable attributes to its outliers. An anomalous point can be measured according to its *a) point difficulty*, *b) semantic variation* and *c) feature relevance*. Although originally the study for these measures was intended to generate anomalies, our study modifies it in order to give quantitative features to an anomaly detection dataset D which can be simply expressed by equation 14:

$$D = \{f_1, f_2\} \quad (14)$$

f_1 corresponds to the ratio of anomalies present in the dataset relative to normal points whereas f_2 corresponds to the dataset's *semantic variation* score. Both these values are constrained to the following:

- 1) $f_1 \leq 0.05$ (5% contamination at most)
- 2) $f_2 \leq 1.5$ (score should be at most 1.5)

1) *Contamination Ratio (f_1):* The contamination ratio is allows a dataset to have the same characterization of anomaly detection situation as seen in literature to express its rarity of occurrence such as those in [9], [4], [5] and [2]. However the datasets used in these studies did not have consistent contamination ratio and only went to what was available in the dataset. In our experiments, given an initial dataset, we sample anomalies to force a ratio of 1%, 2%, 3%, 4% and 5% contamination in order to test the behavior of models under these circumstances. It is important to take note that there is a possibility that the sampled subset could maintain a ratio less than 5% but still have a semantic variation score higher than 1.5. In this case, we do not consider such a subset to be an anomaly detection dataset and resample until both the contamination ratio and semantic variation score is satisfied.

2) *Semantic Variation Score (f_2):* Semantic variation refers to the measure of how an anomalous data point is widely dispersed from the nominal group and fellow outliers. This means that the measure of dispersal should consider both labels of data points in terms of relative distance. Emott et. al chose a random seed point and computed $K - 1$ data points that are closest to it using euclidean distance. This study's approach does not perform any random seeding since we compute a global score for all datapoints within a dataset that has already been subsampled to meet the first constraint of contamination ratio. This is also known as normalized clusterdness measure which can be expressed by equation 15:

$$\log\left(\frac{\sigma_{\text{normal}}^2}{\sigma_{\text{anomaly}}^2}\right) \quad (15)$$

where:

- 1) σ_{normal}^2 is the variance of normal data
- 2) $\sigma_{\text{anomaly}}^2$ is the variance of anomaly data

To deal with multi-dimensional data, we compute for the variance σ^2 of anomaly (or normal) data points X by taking its covariance matrix using equation 16:

TABLE I. INVALID ANOMALY DETECTION DATASET

Dataset	Semantic Variation Score
Iris Versicolor Anomaly	1.63198
Iris Virginica Anomaly	1.57
Iris Setosa Anomaly	1.57212

$$\mathbf{Var}(X) = \mathbf{E}[(X - \mathbf{E}(X))(X - \mathbf{E}(X))^T] \quad (16)$$

We then take trace of the covariance matrix to give us the overall variance using equation 17:

$$\sigma^2 = \text{tr}(\mathbf{Var}(X)) \quad (17)$$

Although in the original paper, it was suggested that such a score did not provide any means of what threshold value constitutes to anomalies (also noting that it was used as a measure for ideal anomaly generation which is a separate study in itself). The data used by most literature suggests that datasets that exhibit an SV score of less than 1.5 tend to be the ones used for benchmarking anomaly detection models thus the constraint was applied together with the context of contamination ratio. An SV score greater than 0 suggests clustered anomalies as opposed to a score less than 0 which suggests more scattered data points. The intuition is that having more clustered anomalies has a higher chance of exhibiting a pattern since it's not simply an incorrect recording of data but also can be repeated with minimal variation within the class. This makes it difficult for density based methods to perform well as clustered points tend to be treated as normal instead of anomaly classes [2].

It is important to note that with these measures and constraints, anomaly detection datasets are not simply defined by the rarity of anomalies that occur but how clustered they are to reflect its difficulty in terms of detection. For example, Table I shows the popular Iris dataset used in [12] is known to be linearly separable but is also treated as an anomaly detection dataset by defining anomalies as part of the tail ends of an interquartile range. If this is the case, we can take a sample of such defined anomalies with a contamination ratio of 5% that exists within the tail ends of interquartile ranges of a class and treat the rest as normal but still have a high semantic variation score. Such conditions will not be treated as anomaly detection benchmarks.

B. Non-Parametric Stochastic Autoencoder Scoring

The proposed method discussed in this research primarily tries to solve for an acceptable t that will yield reliable results for anomaly detection as defined by the expression $f(x) > t$. The method is largely based on training an autoencoder model in order to address non-linear data and providing a new scoring mechanism that takes into account a non-parametric adaptive value assuming no distribution for the data as well as a new loss function that acts as an adaptive regularizer to prevent overfitting. To improve performance, a stochastic process is included and empirically proven to work better compared to using a vanilla autoencoder. Thus, this study names the model the *Non-parametric Stochastic Autoencoder Scoring* model.

1) *Autoencoder Setup*: The first step in the method is to train a standard autoencoder. The topology of the autoencoder will be a shallow one that it will only consist of three layers, the input, the latent and the output layer which has the same dimensionality as the input. The number of latent variables is set to be approximately 3/4 of the original input size. The initial weights of the model are also set symmetrically that is given the initial weights W_{input} connecting the input to the latent layer, the initial weights W_{output} from the latent to the output layer will just be the transpose of W_{input} . Thus, $W_{output} = W_{input}^T$.

For the latent layer, the activation function used was the rectified linear unit (ReLU) given by the equation 18:

$$A_{latent}(x) = \max(x, 0) \quad (18)$$

For the output layer, the activation function used was the sigmoid activation function 19:

$$A_{output}(x) = \frac{1}{1 + \exp(-x)} \quad (19)$$

2) *Autoencoder's Mean Loss*: The new loss function, *Autoencoder's Mean Loss*, to be used for training is a variation of mean squared errors. This loss function is composed of the sum of two terms. The first one is the standard mean squared errors and the second term is the sum of mean squared errors relative to each dimensionality's mean. This is unique to autoencoders as the parameter μ_i can be derived initially from the data set and y_i is simply x_i unlike fully supervised learning methods where y has to be known. This property allows the loss function to adapt to

$$\epsilon(x) = \frac{1}{n} \sum_{i=1}^n (\hat{x}_i - y_i)^2 + \frac{1}{n} \sum_{i=1}^n (\hat{x}_i - \mu_i)^2 \quad (20)$$

A loss function is said to be valid if it is proven to be convex. This is considered to be a valid loss function as proven mathematically since the sum of two convex functions is also convex. Standard back propagation was used for training the autoencoder model.

3) *Stochastic Latent Noise*: A relevant part of the model is adding stochasticity to the latent layer of a trained autoencoder which is referred to as *Stochastic Latent Noise*. The intuition behind this is that if the model is trained only using positive / nominal data points, then determining a threshold to discriminate reconstructed points with high residual errors will still yield to a lot of false negatives. According to the original definition of outliers by Edgeworth, a possible reason for anomalies is the error of observation is the joint result of considerable, but finite, number of small sources of error. This concept is applied to the latent set Z from a trained autoencoder by determining its statistical properties μ_{z_i} and σ_{z_i} and sampling a normal distribution from it with parameter K corresponding to the size of data points found in the distribution. Another parameter d is given referring to the number of dimensions the sampling is applied to. Finally a parameter r is given to represent the ratio of random points taken at the tail end of each distribution to replace the value

at latent index i . Using the trained autoencoder, the decoder part is then ran against these synthetic anomalies to get the reconstructed version at the original dimensional space. These values are then added to the histogram of residual errors to determine t as explained in the next section.

Another way of looking at *Stochastic Latent Noise* is that it's cyclical process of artificial reinforcement learning unique to autoencoders. Traditional machine learning algorithms rely on data to dictate the value of weights whereas reinforcement learning lets the model itself dictate the data. In a similar fashion, the autoencoder first learns of the approximation of the identity function from the data set and based on the weights forces some random aspect to its latent layer representing a compressed version of the data. This is then projected to reconstructed data or randomly synthesized instances that resemble the data as understood by the autoencoder model. An illustration of this can be seen in Fig. 4.

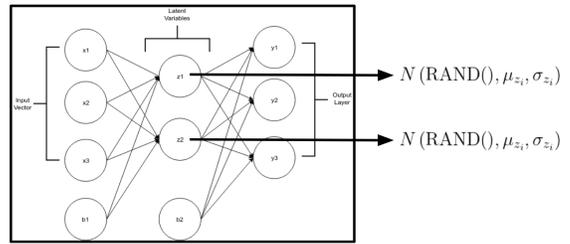


Fig. 4. Stochastic Latent Noise

4) *Determining the Threshold*: The value of t is taken as the midpoint of the identified bin threshold for a histogram of residual errors as built by the reconstructed data points from the trained autoencoder including the added synthetic points from the previous section.

Histogram of Residual Errors: From the set of \hat{X} , the set of residual errors X_ϵ are used to build a histogram of q bins with starting range $\min(X_\epsilon)$ and ending in $\max(X_\epsilon)$. Each bin has its own local minimum and maximum values whose interval is given by equation 21.

$$\text{interval} = \frac{\max(X_\epsilon) - \min(X_\epsilon)}{p} \quad (21)$$

The value of p is automatically set using Freedman-Diaconis rule as seen in equation 22:

$$p = 2 \left(\frac{\text{IQR}(X_\epsilon)}{\sqrt[3]{n}} \right) \quad (22)$$

An example result of the generated histogram of residual errors is seen in Fig. 5.

Determining t From Histogram of Residual Errors: To solve for t , given the histogram of residual errors, we first apply a head-tail break function (HTB) to return an array of possible break points. The following shows an implementation of the HTB algorithm:

Listing 1: HTB Python Implementation

```
def htb(data):
```

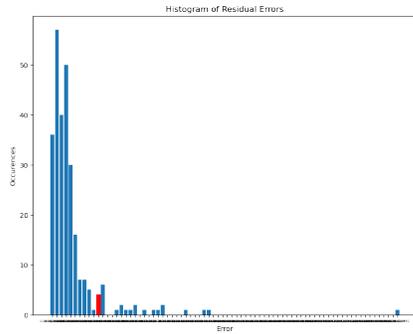


Fig. 5. Histogram of Residual Errors

```

outp = [] # array of break points

def htb_inner(data):
    dl = float(len(data))
    dm = sum(data) / dl
    head = [_ for _ in data if _ > dm]
    outp.append(dm)
    c_head = len(head) > 1
    c_thresh = len(head) / dl < 0.40
    while c_head and c_thresh:
        return htb_inner(head)

htb_inner(data)

return outp

```

For each break point, we get a candidate threshold by passing the bin configuration which is an array of values representing the bins (i.e. bin_0 and bin_1 defines the minimum and maximum value ranges of the first bin) and the occurrence counts in the form also of an array (i.e. in this case, the size of the occurrence counts). The final threshold is then naively selected to be the minimum from the array of candidate thresholds.

The implementation of fetching the threshold from the histogram is given by the following:

Listing 2: Fetch Candidate Threshold

```

def fetch_threshold(bs, counts, bp):
    index = 0
    m = 999999999
    t = -1

    for i in range(len(counts)):
        if abs(counts[i] - bp) <= m:
            m = abs(counts[i] - bp)
            index = i
            l = ((bs[i + 1] - bs[i]) / 2)
            r = bs[i]
            t = l + r

    return t

```

Now that t is determined, to classify an unknown data point x as either an outlier or an anomaly, x is passed to the trained autoencoder and its corresponding x_ϵ is taken. If the value is greater than t , then there is reason to believe that the model had

a hard time reconstructing it and thus flagging it as an anomaly. A visual example of this can be seen in the illustration in Fig. 6.

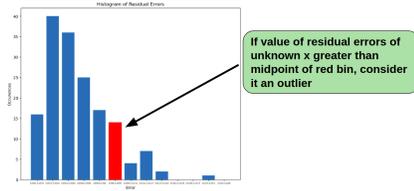


Fig. 6. Threshold Example

C. Methods for Benchmarking

To test the performance of the proposed method against existing, it was compared against 12 standard and state of the art methods for anomaly detection (summarized in Fig. 7 together with year of release) in a semi-supervised fashion where parameters of these methods were manually optimized to get the best possible results. These methods are grouped according to the nature of their methodologies as follows:

Ensemble	Linear	Probabilistic	Proximity
Isolation Forest (2008)	R. Covariance (1994)	ABOD (2008)	LoF (2000)
LODA (2016)	OneClass SVM (2001)	SOS (2012)	CBLOF (2003)
LSC-POE (2019)	MCD (2004)	COPOD (2020)	HBOS (2012)

Fig. 7. Summary of Methods

1) Ensemble:

- 1) Isolation Forest (ISO-F) [3]
- 2) Lightweight On-line Detector of Anomalies (LODA) [13]
- 3) Locally Selective Combination of Parallel Outlier Ensembles (LSCP) [14]

2) Linear:

- 1) Minimum Covariance Determinant (MCD) [5]
- 2) Robust Covariance (ROB-COV) [15]
- 3) OneClass SVM (OC-SVM) [6]

3) Probabilistic:

- 1) Angle-Based Outlier Detection (ABOD) [16]
- 2) Stochastic Outlier Selection (SOS) [17]
- 3) Copula-Based Outlier Detection (COPOD) [18]

4) Proximity:

- 1) Local Outlier Factor (LOF) [4]
- 2) Clustering-Based Local Outlier Factor (CBLOF) [19]
- 3) Histogram-Based Outlier Score (HBOS) [20]

D. Evaluating Performance

To evaluate the results of the proposed method against existing categories of methods mentioned in the previous section, for each sampled dataset (80 datasets in total), 10 simulations were conducted (total of 800 runs: 16 initial datasets

partitioned to five different contamination configurations) with the MCC (Matthew’s Correlation Coefficient) score extracted and applied to a two tailed t-test score with a significance level of 0.05. This allowed the study to determine with confidence if the proposed method is either:

- 1) Significantly better than other methods
- 2) Better but not significantly better than other methods
- 3) Poorer (at least one method is better than the proposed method) but not significantly poorer
- 4) Significantly poorer (at least one method is significantly better than the proposed method)

MCC (Matthew’s Correlation Coefficient) was preferred over the commonly used F1-score due to the mathematical properties mentioned in [21] making it more ideal for anomaly detection with extremely biased data. For each initial dataset, 70% of normal data was randomly sampled and used for training with the remaining 30% used for evaluation. MCC is a measure of correctly predicting both majority of nominal and majority of anomalies. A value of -1 is reached for a perfect misclassification. A value of 1 is reached for a perfect classification. A value of 0 indicates a performance the same as a coin toss. This score can be computed by equation 23

$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP) \cdot (TP + FN) \cdot (TN + FP) \cdot (TN + FN)}} \quad (23)$$

E. Datasets

Given the quantitative measures to define anomaly detection datasets, the following datasets were sampled from 16 datasets to meet the constraints mentioned in order to come up with a total of 80 datasets as shown in Tables II, III, IV, V, VI, VII, VIII, IX, X, XI, XII, XIII, XIV, XV, XVI and XVII. These tables show the values for each dataset’s *categorical measures* in terms of contamination ratio and semantic variation score that characterizes it as an anomaly detection problem which also meets the constraints as mentioned previously.

IV. RESULTS AND ANALYSIS

A. Overview

A summary of the results is illustrated in Fig. 8 where:

- Green cells ● mean that the proposed method did significantly better
- Blue cells ● mean that the proposed method did better but not significantly better
- Yellow cells ● mean that the proposed method did poorer but not significantly poorer
- Red cell ● mean that the proposed method did significantly poorer

To better assess the results of the proposed model, each dataset was treated as a point in two dimensional space where each dimension corresponds to the a dataset’s categorical measure. This allowed us to see if the proposed method performs relatively well in a certain range as bound by the

Dataset	Ensemble					Linear					Probability					Proximity					
	1%	2%	3%	4%	5%	1%	2%	3%	4%	5%	1%	2%	3%	4%	5%	1%	2%	3%	4%	5%	
Anthyroid	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●
Backdoor	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●
Bank	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●
Cover	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●
Credit Card Fraud	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●
Donors	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●
KDDCup99	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●
Magic04	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●
Mammography	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●
Musk	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●
Pageblocks	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●
Seismic	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●
Shuttle	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●
Speech	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●
Synthetic	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●
Waveform	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●

Fig. 8. Overview of Experimental Results

values of categorical measures. Consequently, this allowed us to see at what range the method fails and where a certain category of outlier detection models would prove to be more useful. The next few sections would go through each anomaly detection category and how the proposed method performed comparatively.

B. Performance vs Ensemble Methods

The projected categorical measures can be seen in Fig. 9 where majority of the datasets that the proposed method did significantly better against fall under the range of -0.75 at minimum and 0.60 at maximum in terms of semantic variation score regardless of outlier ratio.

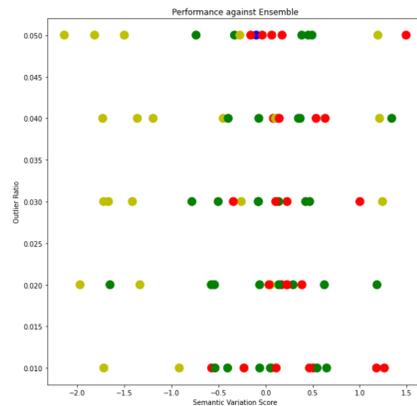


Fig. 9. Performance vs Ensemble

Ensemble methods require calibration of the sub-algorithms used and parameter tuning for each which makes it dependent on which exact algorithms are part of the ensemble. These methods tend to perform better than the proposed method semantic variation scores are negative contrary to the other method categories as seen in the next section. It is still noted though that the proposed method doesn’t require such dependency on other algorithms making it less complex to calibrate.

C. Performance vs Linear Methods

The projected categorical measures can be seen in Fig. 10 where majority of the datasets that the proposed method did significantly better against has a semantic variation score of at most -0.50 regardless of outlier ratio. Although there are some instances beyond -0.50 that the proposed method can

outperform linear methods, in most cases linear methods work significantly better.

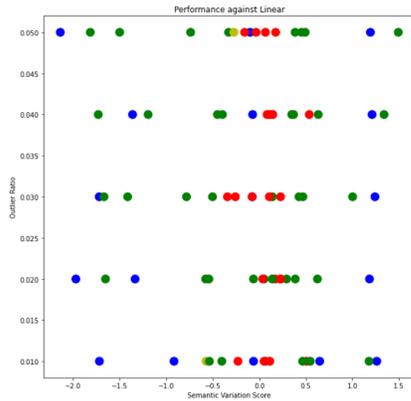


Fig. 10. Performance vs Linear

As with the results of proximity and probability based methods, linear methods fail to outscore the proposed methods in the negatively scored datasets in terms of semantic variation. This suggests that neural network based models can take advantage of non-linear properties that are not otherwise captured by more statistical properties of linear methods. In addition, it suggests that because of the negative values, the anomalies present in such datasets tend to be more scattered (less clustered) with more variation expressing a non-linear behavior.

D. Performance vs Probability Methods

The projected categorical measures can be seen in Fig. 11 where majority of the datasets that the proposed method did significantly better against has a semantic variation score of at most -0.50 regardless of outlier ratio. As with linear methods, compared to probability methods, the method may at times perform better if the semantic variation score is greater than -0.50 but in most cases it performs either poorer or significantly poorer in that domain.

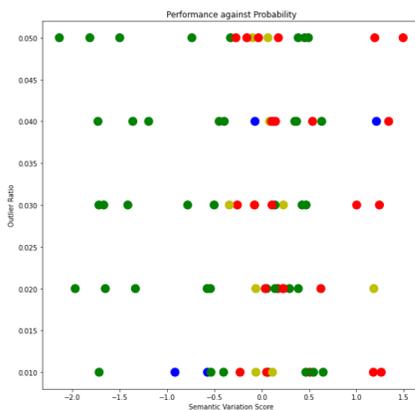


Fig. 11. Performance vs Probability

E. Performance vs Proximity Methods

The projected categorical measures can be seen in Fig. 12 and in similar comparison to linear and probability based

methods, the proposed method performs significantly better than proximity based methods if the semantic variation score of the dataset to be evaluated is less than -0.50 regardless of outlier ratio.

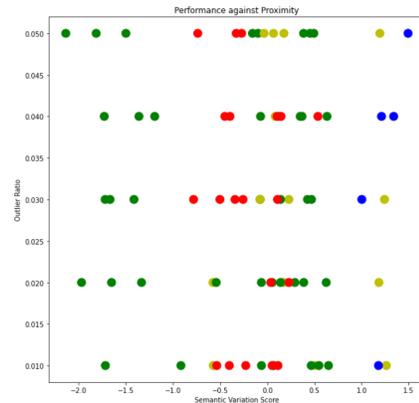


Fig. 12. Performance vs Proximity

As with the previous categories, the proposed method tend to perform better when the anomalies are less clustered. As with the nature of proximity based methods which is largely based on distance metrics such as nearest neighbors, a large variance of anomalies present will tend to fail as seen in the datasets that are negatively scored in terms of semantic variation.

F. General Discussion

As a general statement, it can be seen that in terms of the MCC performance against datasets constrained with the categorical measures, the proposed model can perform significantly well regardless of outlier ratio in most cases if the semantic variation score leans towards negative values, specifically -0.05 . As mentioned by Emott et. al. [2], a dataset quantitatively characterized with a negative semantic variation score suggests that anomalies are more scattered in nature due to having a higher value in terms of variance that exceeds that of the variance of normal data points (taking the log of a value less than 1 as computed by semantic variation, will yield a negative value as a result of the variance of anomalies is higher than that of normal points).

Restricting the contamination ratio to at most 5% keeps it aligned with the concept of rarity as what most literature in anomaly detection states. It is important to note however that not all the initial datasets (that is, the dataset's original form without the sampled subsets for evaluation with applied categorical measures) necessarily comply with the constraints of categorical measures. Violating these constraints would not constitute to proper anomaly detection studies as it will have a positive semantic variation score (suggested threshold of the study is 1.5) that suggests occurrences of anomalies would tend to form certain patterns which does not appropriately reflect the description of anomalies in literature (incorrect data production and rare occurrence) and therefore could be a means of it being treated as either a biased two class classification problem wherein the minority class is not considered an anomaly.

In terms of application, if a scenario in the real world can determine anomalies that tend to cluster in such a way,

then the proposed method is ideal for it in a sense that it will allow researchers / practitioners / stakeholders to take advantage of not needing to manually adjust inferencing parameters (reconstruction error threshold) upon usage. Being a neural network based model, it does seem to validate the approximation of non-linear functions that tend to address the unpredictable behavior of occurrences of anomalies. On the flip side however, anomalies that tend to be more clustered together thus exhibiting a higher semantic variation score (greater than -0.50 or a positive value for that matter), would still fall under the notion that the effectiveness of an anomaly detection algorithm will still be on a case to case basis.

Applying categorical measures allowed us to quantify the nature of an anomaly detection dataset and observe the behavior of various methods in terms of its MCC performance. With these measures, we can derive insights as to which possible range of values certain methods can work well against as opposed to the general characteristics of the anomaly detection problem where anomalies are simply said to rarely occur or fall within a certain deviation from what is normal. Such deviation can't be quantified as anomalies themselves are subject to the domain where data is captured or produced thereby not allowing a standard objective definition of it. With categorical measures, we can at the very least, quantify the structure of the dataset in relation to the existence of anomalies and its relation to normal data points.

V. CONCLUSION

The effectiveness of classification methods for anomaly detection are traditionally dependent on looking for the best parameters that fits a certain scenario or dataset at hand. This is primarily due to the fact that anomalies themselves cannot be quantified or given objective characteristics for all domains. Even in current literature, anomalies are restricted to definitions that vary from scenario to scenario in terms of rarity and what value to be used to express magnitude of deviation from what is normal. Given these, this study pushes the definition further by providing a quantitative definition not to the anomalies themselves but in context of datasets that are considered to be an anomaly detection problem so as to give insights as to how well anomaly detection methods perform. The study refers to these as *categorical measures*.

Anomaly detection datasets (specifically point anomalies and not context or time series type) characterized with *categorical measures* in this study was constrained to the following conditions:

- 1) The point anomaly instances comprise of at most 5% of the evaluated dataset to ensure its rarity of occurrence.
- 2) The dataset's semantic variation score does not exceed 1.5, since higher scores generally imply that there is clustering among the point anomaly instances, and this may indicate the presence of a non-anomalous process that generated the "anomaly" instances.

Both conditions have to be satisfied before a dataset can be considered as part of an anomaly detection study. Violating these constraints would not constitute to proper anomaly detection as the score suggests certain clusters forming depicting

a pattern where methods that are density based are more likely to fail. Since the first constraint expresses rarity, it is still a ratio and one can still derive a subsample that meets the first criteria but fails in the second. Therefore, both have to be met to constitute to a valid anomaly detection problem instead of just leaning on the notion of rarity.

With the proposed method with a neural network autoencoder model as its base, it can be modified to be adaptive in terms of automatically setting the parameter in the inferencing stage that allows scoring of anomalies to not be dependent on prior information such as assumed distribution on the data itself or domain expert. This was done by allowing the discovery of the threshold value, a parameter that has traditionally been set manually by neural network based models, to be naturally formed by the distribution of reconstruction errors which assumes to exhibit a long tail distribution. Compared to existing methods that have been tested throughout this study, the proposed method performed comparatively well in an identified domain of negatively scored semantic variation value of anomaly detection datasets suggesting that it works well in scenarios with more variation in the anomalies present. This has been proven through experimentation on the MCC metric where in most cases, the proposed method performed significantly better if the dataset has a score of -0.5 or less in terms of semantic variation even when parameters of the existing methods have been optimized whereas the proposed method had its parameter configured automatically all throughout. Apart from the family of ensemble method where the performance of the proposed method works better in the range of -0.75 to 0.60 , as well as the positively scored datasets, it still remains a case to case basis as other methods work better or worse than the proposed method without any evident generalization. This could be up for investigation in future work.

Anomaly detection in general is still an open problem without any standard objective definition. But with the case of this study, the narrative can be progressed towards examining behavior of methods given categorical measures of the datasets themselves to constitute to a working quantitative definition of an anomaly detection problem.

ACKNOWLEDGMENT

The authors would like to thank the Ateneo de Manila University and the University Research Council for supporting this study.

REFERENCES

- [1] M. Salehi and L. Rashidi, "A survey on anomaly detection in evolving data: [with application to forest fire risk prediction]," *SIGKDD Explor. Newsl.*, vol. 20, no. 1, pp. 13–23, May 2018. [Online]. Available: <http://doi.acm.org/10.1145/3229329.3229332>
- [2] A. Emmott, S. Das, T. Dietterich, A. Fern, and W.-K. Wong, "Systematic construction of anomaly detection benchmarks from real data," 08 2013, pp. 16–21.
- [3] F. T. Liu, K. M. Ting, and Z.-H. Zhou, "Isolation forest," in *Proceedings of the 2008 Eighth IEEE International Conference on Data Mining*, ser. ICDM '08. USA: IEEE Computer Society, 2008, p. 413–422. [Online]. Available: <https://doi.org/10.1109/ICDM.2008.17>
- [4] M. Breunig, H.-P. Kriegel, R. Ng, and J. Sander, "Lof: Identifying density-based local outliers." vol. 29, 06 2000, pp. 93–104.

[5] P. Rousseeuw and K. Driessen, "A fast algorithm for the minimum covariance determinant estimator," *Technometrics*, vol. 41, pp. 212–223, 08 1999.

[6] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, Sep 1995. [Online]. Available: <https://doi.org/10.1007/BF00994018>

[7] B. Schölkopf, R. Williamson, A. Smola, J. Shawe-Taylor, and J. Platt, "Support vector method for novelty detection," vol. 12, 01 1999, pp. 582–588.

[8] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," *ACM Comput. Surv.*, vol. 41, 07 2009.

[9] Y. Ishii and M. Takashi, "Low-cost unsupervised outlier detection by autoencoders with robust estimation," *Journal of Information Processing*, vol. 27, pp. 335–339, 01 2019.

[10] R. Chalapathy, A. Menon, and S. Chawla, "Anomaly detection using one-class neural networks," 02 2018.

[11] T. Amarbayasgalan, B. Jargalsaikhan, and K. Ryu, "Unsupervised novelty detection using deep autoencoders with density based clustering," *Applied Sciences*, vol. 8, 08 2018.

[12] E. Acuna and C. Rodriguez, "An empirical study of the effect of outliers on the misclassification error rate," 11 2004.

[13] T. Pevný, "Loda: Lightweight on-line detector of anomalies," *Mach. Learn.*, vol. 102, no. 2, p. 275–304, feb 2016. [Online]. Available: <https://doi.org/10.1007/s10994-015-5521-0>

[14] Y. Zhao, Z. Nasrullah, M. Hryniewicki, and Z. Li, "Lscp: Locally selective combination in parallel outlier ensembles," 01 2019.

[15] A. Stromberg, "Robust covariance estimates based on resampling," pp. 321–334, 02 1997. [Online]. Available: [https://doi.org/10.1016/S0378-3758\(96\)00051-1](https://doi.org/10.1016/S0378-3758(96)00051-1)

[16] X. Li, J. C. Lv, and D. Cheng, "Angle-based outlier detection algorithm with more stable relationships," in *Proceedings of the 18th Asia Pacific Symposium on Intelligent and Evolutionary Systems, Volume 1*, H. Handa, H. Ishibuchi, Y.-S. Ong, and K. C. Tan, Eds. Cham: Springer International Publishing, 2015, pp. 433–446.

[17] J. Janssens, "Outlier selection and one-class classification," Ph.D. dissertation, 2013, series: TiCC Ph.D. Series Volume: 27.

[18] Z. Li, Y. Zhao, N. Botta, C. Ionescu, and X. Hu, "COPOD: Copula-based outlier detection," in *2020 IEEE International Conference on Data Mining (ICDM)*. IEEE, nov 2020. [Online]. Available: <https://doi.org/10.1109/2Ficdm50108.2020.00135>

[19] Z. He, X. Xu, and S. Deng, "Discovering cluster-based local outliers," *Pattern Recognition Letters*, vol. 24, no. 9, pp. 1641–1650, 2003. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0167865503000035>

[20] M. Goldstein and A. Dengel, "Histogram-based outlier score (hbos): A fast unsupervised anomaly detection algorithm," 09 2012.

[21] D. Chicco and G. Jurman, "The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation," *BMC Genomics*, vol. 21, 12 2020.

APPENDIX

TABLE II. ANNTHYROID DATASETS

Dataset	Contamination Ratio	Semantic Variation Score
Annthyroid1	1%	0.508125181
Annthyroid2	2%	0.1707903902
Annthyroid3	3%	0.2287771725
Annthyroid4	4%	0.08389630917
Annthyroid5	5%	0.06585469485

TABLE III. BACKDOOR DATASETS

Dataset	Contamination Ratio	Semantic Variation Score
Backdoor1	1%	0.5475187707
Backdoor2	2%	0.1404898671
Backdoor3	3%	0.1396325904
Backdoor4	4%	0.3682554791
Backdoor5	5%	0.4539561087

TABLE IV. BANK NOTES DATASETS

Dataset	Contamination Ratio	Semantic Variation Score
Bank1	1%	-0.06208764389
Bank2	2%	-0.06267816812
Bank3	3%	-0.07630856422
Bank4	4%	-0.07151140273
Bank5	5%	-0.09810198756

TABLE V. COVER DATASETS

Dataset	Contamination Ratio	Semantic Variation Score
Cover1	1%	0.46640986
Cover2	2%	0.3869515162
Cover3	3%	0.4684336152
Cover4	4%	0.6339205037
Cover5	5%	0.4937923986

TABLE VI. CREDIT CARD FRAUD DATASETS

Dataset	Contamination Ratio	Semantic Variation Score
CCF1	1%	-1.71859332
CCF2	2%	-1.971678786
CCF3	3%	-1.721024117
CCF4	4%	-1.36247447
CCF5	5%	-2.139169117

TABLE VII. DONORS DATASETS

Dataset	Contamination Ratio	Semantic Variation Score
Donors1	1%	1.262872064
Donors2	2%	1.184351675
Donors3	3%	1.243335783
Donors4	4%	1.212204939
Donors5	5%	1.194491206

TABLE VIII. KDDCUP99 DATASETS

Dataset	Contamination Ratio	Semantic Variation Score
KDDCup991	1%	0.06888742085
KDDCup992	2%	-1.653239859
KDDCup993	3%	-1.669393574
KDDCup994	4%	-1.730795094
KDDCup995	5%	-1.501027102

TABLE IX. MAGIC04 DATASETS

Dataset	Contamination Ratio	Semantic Variation Score
Magic041	1%	-0.4032003924
Magic042	2%	-0.5770861584
Magic043	3%	-0.5031427544
Magic044	4%	-0.4501034267
Magic045	5%	-0.3297699596

TABLE X. MAGIC04 DATASETS

Dataset	Contamination Ratio	Semantic Variation Score
Mammography1	1%	1.179545636
Mammography2	2%	0.6250250795
Mammography3	3%	1.003284824
Mammography4	4%	1.341220359
Mammography5	5%	1.493847255

TABLE XI. MUSK DATASETS

Dataset	Contamination Ratio	Semantic Variation Score
Musk1	1%	0.6492543029
Musk2	2%	0.2927728505
Musk3	3%	0.4260056274
Musk4	4%	0.3493787179
Musk5	5%	0.3849209583

TABLE XII. PAGEBLOCKS DATASETS

Dataset	Contamination Ratio	Semantic Variation Score
Pageblocks1	1%	-0.9167593238
Pageblocks2	2%	-1.335170795
Pageblocks3	3%	-1.41523494
Pageblocks4	4%	-1.1950763
Pageblocks5	5%	-1.816508555

TABLE XIII. SEISMIC DATASETS

Dataset	Contamination Ratio	Semantic Variation Score
Seismic1	1%	-0.5731575761
Seismic2	2%	0.2270706794
Seismic3	3%	-0.07664012976
Seismic4	4%	0.1222953365
Seismic5	5%	-0.1572111807

TABLE XIV. SHUTTLE DATASETS

Dataset	Contamination Ratio	Semantic Variation Score
Shuttle1	1%	-0.5363501714
Shuttle2	2%	-0.5423016379
Shuttle3	3%	-0.7827652121
Shuttle4	4%	-0.3966784238
Shuttle5	5%	-0.7382668293

TABLE XV. SPEECH DATASETS

Dataset	Contamination Ratio	Semantic Variation Score
Speech1	1%	0.05144099108
Speech2	2%	0.05144099108
Speech3	3%	-0.2575904149
Speech4	4%	0.1041183734
Speech5	5%	-0.2727381015

TABLE XVI. SYNTHETIC DATASETS

Dataset	Contamination Ratio	Semantic Variation Score
Synthetic1	1%	-0.2289063831
Synthetic2	2%	0.03732289785
Synthetic3	3%	0.1082732279
Synthetic4	4%	0.1454660726
Synthetic5	5%	-0.03449716229

TABLE XVII. WAVEFORM DATASETS

Dataset	Contamination Ratio	Semantic Variation Score
Waveform1	1%	0.1132031272
Waveform2	2%	0.2289010261
Waveform3	3%	-0.3424085047
Waveform4	4%	0.5371335689
Waveform5	5%	0.1755232835

COVID-19 Cases Detection from Chest X-Ray Images using CNN based Deep Learning Model

Md Amirul Islam¹, Giovanni Stea², Sultan Mahmud³, Kh. Mustafizur Rahman⁴
Department of Information Engineering, University of Pisa, Italy^{1,2,3}
Faculty of Science & Technology, Bangladesh University of Professionals, Bangladesh⁴

Abstract—COVID-19 has recently manifested as one of the most serious life-threatening infections and is still circulating globally. COVID-19 can be contained to a considerable extent if a patient can know their COVID-19 infection at a possible earlier time, and they can be isolated from other individuals. Recently, researchers have explored AI (Artificial Intelligence) based technologies like deep learning and machine learning strategies to identify COVID-19 infection. Individuals can detect COVID-19 disease using their phones or computers, dispensing with the need for clinical specimens or visits to a diagnostic center. This can significantly reduce the risk of spreading COVID-19 farther from a probably infected patient. Motivated by the above, we propose a deep-learning model using CNN (Convolutional Neural Networks) to autonomously diagnose COVID-19 disease from CXR (Chest X-ray) images. The dataset used to train our model includes 10293 X-ray images, with 875 X-ray images from COVID-19 cases. The dataset contains three different classes of the tuple: COVID-19, pneumonia, and normal cases. The empirical outcomes show that the proposed model achieved 97% specificity, 96.3% accuracy, 96% precision, 96% sensitivity, and 96% F1-score, respectively, which are better than the available works, despite using a CNN with fewer layers than those.

Keywords—COVID-19; CNN; deep learning; machine learning; chest X-ray

I. INTRODUCTION

Wuhan, a business hub in China's Hubei province, saw a fresh coronavirus outbreak in 2019. Researchers in China termed the novel virus the 2019 n-Cov or the Wuhan virus [1]. Officially a study group found this virus, and they named it SARS-CoV-2. During the coronavirus infection outbreak in 2019, the international committee was labeled coronavirus disease 19 (COVID-19) [2, 3, 4]. In the first place, coronaviruses have caused disease in people who have been exposed to wild animals, primarily bats and rats [5, 6, 7]. After almost three years since its first appearance, COVID-19 shows no signs of abating, and has already infected half a billion individuals and claimed more than six million deaths.

COVID-19 is a disease caused by a virus source that aggravates the lungs, causing pneumonia in patients. However, the treatment and medication of these cases are quite different from pneumonia cases arising from other viruses or bacteria. If an individual shows symptoms of COVID-19, certain precautionary steps are taken in conjunction with the diagnosis. COVID-19 patient is isolated for a certain number of days in order to contain the further spread of the infection. Therefore, to stop the spread of COVID-19, accurate and timely identification of pneumonia caused by the virus is a critical issue.

The WHO approved a testing method based on Polymerase chain reaction (RT-PCR), whereby short DNA or RNA sequences are analyzed and replicated or intensified [8]. In some circumstances, more than one test may be required to rule out the possibility of coronavirus infection. According to the WHO laboratory research, negative results do not always mean that a person is not infected with COVID-19 [9].

Although RT-PCR examination is the most trustworthy technique to diagnose COVID [10], it is a laborious, time consuming, complex manual procedure, with kits in short supply depending on the time and place. Furthermore, the test is unpleasant and slightly invasive, as it entails collecting nasopharyngeal swabs. COVID-19 attacks the human airway epithelial (hAE) cells. Accordingly, X-rays can be used as a potential specimen to know the damaged portions of a COVID-19 patient's lungs. Detection of COVID-19 using X-ray images of a patient might be a helpful technique due to its fast speed, contained cost, and wide range of applications [11].

The insufficient supply of COVID-19 scanning workstations and research kits poses challenges for the medical practitioners and personnel to cope up with COVID-19. Rapid and effective identification of suspicious COVID-19 cases is of fundamental importance to contain the spread of the infection. The need for repeated checks to ascertain the actual condition and make effective decisions is often exponential growth in critical situations for patients. Therefore, we set the following objectives to design a machine-learning based COVID-19 case detection system.

We summarize the objectives as follows:

- Assisting radiologists and medical specialists in detecting subtle and gradual changes in X-rays that might otherwise go unnoticed;
- Many people in developing countries do not have access to radiologists due to high costs. This tool could help them read their X-ray images to classify them as COVID-19, pneumonia, and normal;
- Creating a model to scan complex data such as CT and MRI images to detect COVID-19 cases.

A deep-learning approach, such as CNNs, is able to learn and deduce properties of a very complex nonlinear functions autonomously, without the need for human intervention. CNNs act exclusively on input data after a supervised training phase. For example, the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) reported a well-known scenario when a

model outperformed humans in the task of classifying images in 2015 [12].

The primary purpose of this study is to build a deep-learning paradigm to diagnose pneumonia from CXR images as well as the location and positioning of abnormalities in X-rays. In addition, we propose a deep-learning-based model to detect cases of COVID-19, influenza, and normal X-ray picture classification. This model provides a low-cost method for radiologists and medical professionals to cross-check their interpretations and recognize any possible results that might otherwise have been overlooked. The model is trained with the dataset collected from Kaggle and GitHub. The radiological society and researchers recently formed these datasets from patients' X-ray images during the COVID-19 outbreak. To improve the model's accuracy, we adopt clinical image engineering that has consolidated imaging computing approaches such as pre-processing techniques, classification approaches, and illness screening and localization in a sequential manner.

Following are the contributions of this paper:

- We developed a CNN-based architecture to detect COVID-19 from patients' X-rays after collecting and pre-processing X-ray data. We trained our model by using Keras on top of TensorFlow. The proposed approach achieves 96.3% accuracy for the Pneumonia, Normal, and COVID-19 classes. Despite our model being less complex (as measured by the number of layers in the CNN) than the available ones, it outperforms them all in accuracy.
- Grad-CAM has been employed to detect characteristic features from X-ray images to aid visually interpretative decision-making about COVID-19.

The rest of this paper is summarized as follows: In Section II, recent research works are represented. Our suggested approach and material with CNN-based architecture for the COVID-19 detection system are discussed in Section III. Experimental results and discussion are illustrated in Section IV and Section V concludes our paper and highlights directions for future work. Table I reports a list of the acronyms used in this paper.

II. LITERATURE REVIEW

Researchers have been utilizing different deep-learning strategies to identify COVID-19 from clinical images such as X-rays and CT scans of the chest. Rahimzadeh et al. [13] built a joint CNN-assisted ResNet50V2 and Xception model for categorizing COVID-19 cases by employing CXR images. The dataset fed into this method included 8851 images of healthy people, 6054 images of pneumonia patients, and 80 images of COVID-19 patients. In this model, 633 images were chosen for each of the eight training phases. For COVID-19 cases, this system achieved an experimental result of 99.56% accuracy and 80.53% recall.

Alqudah et al. [14] utilized AI strategies to implement a method for detecting COVID-19 from CXR images. Several machine-learning strategies like Support Vector Machine (SVM) and Random Forest (RF) were picked for classification purposes. The system achieved 95.2% accuracy, 93.3% sensitivity, and 100% specificity. GAN with deep learning

to identify the COVID-19 from CXR images was suggested by Loey et al. [7]. GoogleNet, AlexNet, and ResNet18 are the pre-trained models used by this proposed framework for identifying COVID-19 cases. The GoogleNet pre-trained model was picked as a primary deep-learning technique. The model showed test accuracy of 80.6% within the four classes continuity (79% normal cases, 79% bacterial pneumonia cases, 79% images of pneumonia cases, and 69% COVID-19 cases). In the same way, for three classes continuity, AlexNet gained test accuracy 85.2%, and GoogleNet attained test accuracy of 99.9% for the two classes continuity.

Kumar et al. [15] suggested a deep-learning procedure to identify and classify COVID-19 infected patients. The proposed approach was trained by using nine pre-trained processes to extract features from the CXR images, and for the classification they used SVM. The datasets included 158 CXR images of both COVID-19 and non-COVID-19 cases. Obtaining 95.52% F1-score and 95.38% accuracy, ResNet50 combined with the SVM method was a statistically outstanding technique.

Horry et al. [16] suggested a COVID-19 identifying framework from CXRs considering pre-trained methods. The recommended framework applied Inception, ResNet, Xception, and VGG to classify COVID-19 cases. The dataset employed within the framework encompasses 60361 images of normal cases, 322 images of pneumonia cases, and 115 images of COVID-19 cases. They were found recall and precision values 80% for the VGG19 and VGG16 classifiers.

Ucar et al. [17] introduced a COVID-19 identifying framework by utilizing X-ray photos and deep learning techniques. The data set used for the framework contained 1583 photos of normal cases, 76 photos of COVID-19 cases, and 4290 photos of pneumonia cases, and for COVID-19 cases, the system attained 98.3% accuracy. A transfer-learning technique with the CNN is recommended by Apostolopoulos et al. [18]. By extracting essential attributes from CXR images, this technique could instinctively diagnose COVID-19 patients. This recommended framework used five CNN models, including InceptionResNetV2, MoblieNet, Inception, Xception, and VGG19, to sort the COVID-19 CXR images. The dataset for three classes contained 504 CXR images of normal cases, 700 CXR images of pneumonia cases, and 224 CXR images of COVID-19 cases. VGG19 was chosen as the primary method with 98.75% sensitivity, 92.85% specificity, and 93.48% accuracy within the introduced framework.

A complete unique model introduced by Bandyopadhyay et al. [19] that made use of the LSTM-GRU to categorize confirmed automatically, negative, positive, recovered, and death cases of coronavirus. The above framework attained 67.8% accuracy for negative cases, 62% accuracy for death cases, 40.5% accuracy for released cases, and confirmed cases with 87% accuracy.

A deep-learning network to identify COVID-19 cases autonomously from CXRs was presented by Khan et al. [20]. The dataset includes 310 images of normal cases, 330 images of bacterial pneumonia cases, 327 images of the viral pneumonia cases, and 284 images of COVID-19 cases. With COVID-19, the recommended model acquired 93.5% accuracy, 100% recall, and 97% precision. A complete unique framework

TABLE I. LIST OF ACRONYMS USED IN THE PAPER

Notation	Definition
AI	Artificial intelligence
AP	Attending Physician
AUC	Area Under the Curve
COVID-19	Coronavirus
CNN	Convolutional Neural Network
CT	Computed Tomography
CXR	Chest X-ray
CMC	Composite Monte-Carlo
CSL	Cost Sensitive Learning
DNN	Deep Neural Network
FCL	Fully Connected Layers
FP	False Positive
FN	False Negative
Grad-CAM	Gradient Class Activation Map
GAN	Generative Adversarial Network
Grad-CAM	Gradient-Weighted Classes Activation Mapping
GPU	Graphical Processing Unit
ILSVRC	ImageNet Large-Scale Visual Recognition Challenge
KTD	Knowledge Transfer and Distribution
LR	Learning Rate
MRI	Magnetic Resonance Imaging
PCA	Principal Component Analysis
RT-PCR	Reverse Transcription Polymerase Chain Reaction
RF	Random Forest
ReLU	Rectified Linear Unit
ROC	Receiver Operating Characteristic Curve
SARS-CoV-2	CoronaVirus-2 Severe Acute Respiratory Syndrome
SVM	Support Vector Machine
TP	True Positive
TN	True Negative
WHO	World Health Organisation
2019 n-Cov	2019-Novel-Coronavirus

based on a deep-learning approach to identify COVID-19 from CXR images called COVIDX-Net was represented by Hemdan et al. [21]. Authors used eight pre-trained models contained Xception, InceptionResNetV2, InceptionV3, VGG19, DenseNet201, MobileNetV2, and ResNetV2 for their frameworks. The dataset contained 50 CXR images, with 25 images of non-COVID-19 cases and 25 of COVID-19 cases. VGG19 and DenseNet201 achieved 83% precision and 90% accuracy.

In [22], Singh et al. utilized a VGG16 model as a deep transfer-learning framework to identify COVID-19 patients from CT scan images. In this proposed framework, authors used four classifiers models to classify COVID-19 cases, and the PCA extracted the features from CXR images. With the help of the SVM classifier and bagging ensemble method, 95.7% accuracy, 95.3% F1-score, and 95.8% precision achieved effective result.

In [23] the authors presented a lightweight DNN-based mobile app named COVID-MobileXpert, that can use noisy photos of CXRs for point-of-care COVID-19 screening. In addition, they created a new 3-players Knowledge Transfer and Distribution (KTD) framework that includes a pre-trained attending physician network that extracted photo attributes from a gigantic set of lung infection CXR photos. Furthermore, a fine-tuned RF network retains the fundamental CXR imaging attributes to distinguish COVID-19 cases from normal and pneumonia CXR images.

To detect COVID-19 symptoms, Ahuja et al. [24] suggested a transference learning strategy employing a three-phase approach. With a view to indicating the abnormality of the image, several pre-trained models had been applied with the augmented image using ResNet18 deep-learning model, and thus, 99.4% of test cases accuracy were obtained. In [25] the authors

talked about deep learning emerged case study utilizing CMC & fuzzy rule induction to deal with the delimited information for forecasting strategies. To identify COVID-19 cases, Ashraf et al. [26] presented a framework called COVID-CAPS. In [27] authors introduced a deep-learning algorithm, which can perform fast detection of COVID-19 cases, and their system attained 95% accuracy. Karthik et al. [28] suggested a custom CNN-based architecture to detect COVID-19 cases, and for each kind of pneumonia, it can learn unique convolutional filter patterns.

TABLE II. NUMBER OF CNN LAYERS IN THE REVIEWED MODELS

Year	Ref.	Number of Layers
2020	[13]	6
2020	[14]	8
2020	[15]	≥ 1000
2020	[16]	≥ 200
2020	[17]	5 \times 15 Layers/Sublayers
2020	[18]	≥ 1000
2020	[20]	74
2020	[21]	≥ 400
2020	[22]	2500 Hidden Layers with 8 Different Layers
2020	[23]	9
2021	[24]	71
2020	[26]	7
2020	[27]	27 \times 19 Layers/Sublayers
2021	[28]	10
2022	Proposed Model	5

Aside from that, researchers continually contribute to inventing feasible methods to identify COVID-19 cases from CXR images. The above review establishes that deep learning, especially CNN, plays an important role in COVID-19 diseases

detection and classification in medical imaging. However, the recent study has limitations in identifying the COVID-19 virus in normal cases from the CXR images. Our proposed model is novel and different from the others presented above. It has fewer layers, which entails smaller complexity. The number of layers used in the methods reviewed in this section is shown in Table II, and is always larger, and often considerably so. As we will show later, despite being less complex, our model achieves better performance in most metrics.

III. PROPOSED COVID-19 DETECTION SYSTEM

This section presents the primary contribution of the paper: the architectural design and implementation of the suggested methodology.

A. Methodology

The architecture of our system for COVID-19 detection consists of several building blocks as illustrated in Fig. 1.

Initially, raw CXR images are provided in the pre-processing pipeline for performing pre-processing tasks like resizing, normalization, flipping, zooming, and rotation. After the pre-processing phase, the data set is divided into a training and a test set. Next, the proposed system is trained using the training data. Training and validation accuracy and loss are determined after every epoch. The following evaluation metrics have been adopted for performance measurement: accuracy, sensitivity, specificity, F1-score, precision, Area Under the Curve (AUC) using Receiver Operating Characteristics (ROC) and confusion matrix.

1) *Collection of Dataset and Explanation:* As COVID-19 has recently broken out, datasets with more sample X-rays tagged with COVID-19 cases are not available yet. As a result, collecting data on various image sources of normal, COVID-19, and pneumonia cases need to be collected. For COVID-19 cases, 2875 CXR images have been collected from GitHub [29] and Kaggle [30]. For pneumonia and normal cases, 4200 and 3218 CXR images are collected from the Kaggle repository [30, 31]. In total, our dataset contains 10293 CXR images. Later, the images are resized to a resolution of 224x224 pixels. Fig. 2 depicts the visualization of several CXR images of every class.

The number of CXR images of each set was partitioned in Table III

TABLE III. DATASET

Normal	Covid	Pneumonia	Total
3218	2875	4200	10293

2) *Image Pre-Processing:* To improve the model's efficiency, pre-processing of the image is required. In this work, we also carried out certain pre-processing activities to produce better performance. The pre-processing techniques used in this research are listed in this section.

A.1 Resize Picture to Capture the Central Portion and Remove Black Bars: Typically, square images are required to train CNN models. The CXR data obtained from various sources differs in sizes. So, in order to feed CXR images into

the model, we need to scale the images down to a square shape. However, this causes an asymmetry of the images as seen in the Fig. 3.

To tackle the distortion of the input images, we applied Pasa et al. [32]'s technique to remove the central area and delete black bars. Fig. 4 depicts a pre-processed image obtained by carrying out the following operations:

- 1) If any black bands appear at the image's margins, they are discarded;
- 2) The image's dimension is warped until the minimum boundary counts 224 pixels;
- 3) Retrieve the 224×224 pixel core area.

A.2 Normalization: Subsequently, images are normalized and transformed to the proper data type. Original images are grayscale with individual pixels coded as Unit8 type with values in range 0-255. For the Keras model, data is provided in float32 type. Therefore, the pre-processed images have to be converted into float32, normalized to a range [0; 1]. To normalize every pixel, the data value is divided by the highest value of uint8, which is 255.

A.3 Data Augmentation: Deep-learning methods, such as CNN, generates more accurate outcomes if larger numbers of full-sized images can be used. As a result, data augmentation is useful in creating the first training images that are backed up by extra data. We conducted the subsequent data augmentation procedures during training:

- 5 to 10 degrees of spontaneous rotation;
- Zooming within a range of +10% and -10%;
- Flipping horizontally.

3) *Proposed CNN Model Development:* A deep-learning model has been built in this paper to diagnose COVID-19 incidents. The model is trained using a dataset having three types of CXR images. The suggested CNN model structure for COVID-19 detection is shown in Fig. 5.

The proposed CNN model has five convolutional blocks. Each convolutional block has multiple layers, and each layer has one activation function named Rectified Linear Unit (ReLU). The third block and fourth block have a dropout layer to reduce the over-fitting problem. Two fully connected layers (FCLs) have been used, the first FCL is used with the dropout layer, and the last FCL is connected with the softmax classifier.

A.1 Convolution Layer: Every convolution block of this model has multiple convolution layers. The first convolution block consists of conv1-layer1, and conv1-layer2. Similarly, second convolution block has conv2-layer1, and conv2-layer2. Third convolution block has conv3-layer1, conv3-layer2. Fourth convolution block consists of conv4-layer1, and conv4-layer2, and final fifth block has conv5-layer1, and conv5-layer2. The fourth block and fifth block have an extra dropout layer to reduce the overfitting problem. Conv1 block has used a total of 32 filters where every filter size is 5×5 . Conv2 block uses a total of 64 filters where every filter size is 3×3 . Conv3 block uses a total of 128 filters where every filter size is 3×3 . Conv4 block uses a total of 256 filters where every filter size is 3×3 , and similarly, Conv5 block has used a total of 512 filters where every filter size is 3×3 .

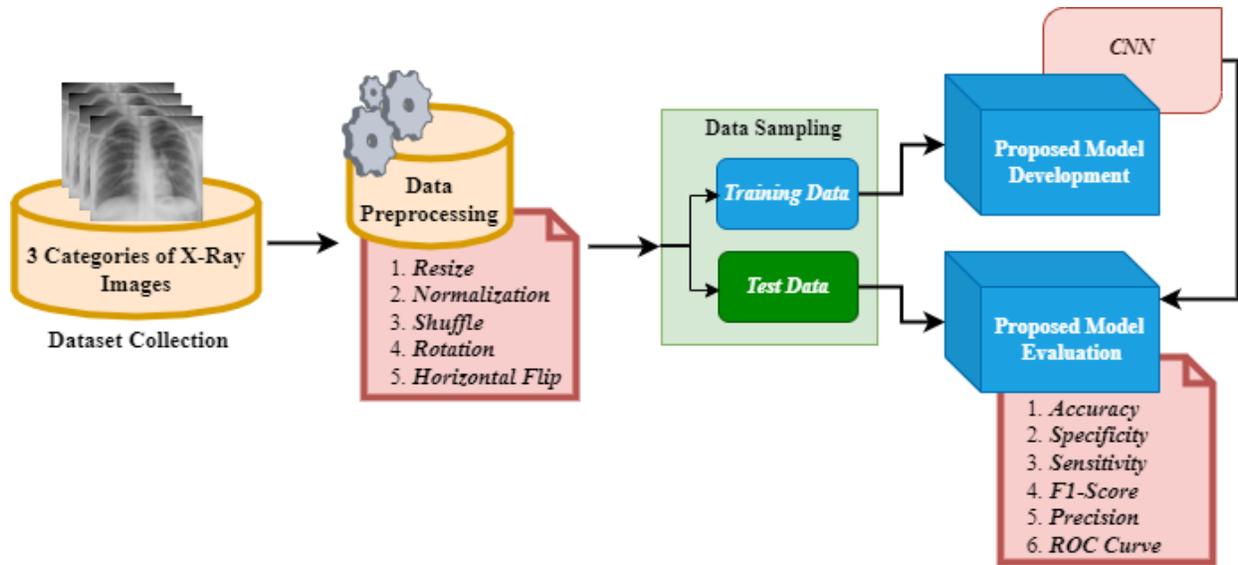


Fig. 1. Proposed Architecture for COVID-19 Detection System.

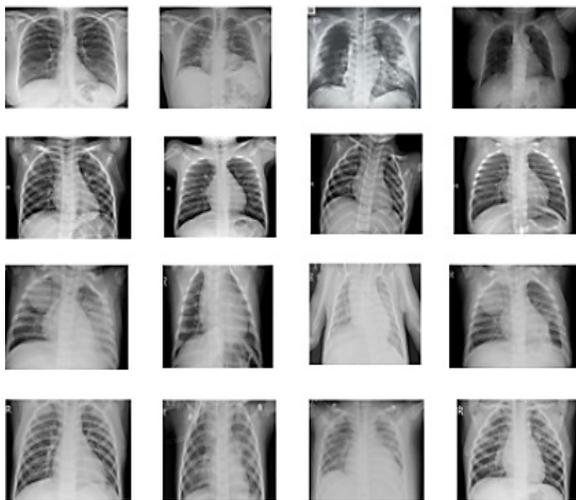


Fig. 2. X-ray Images

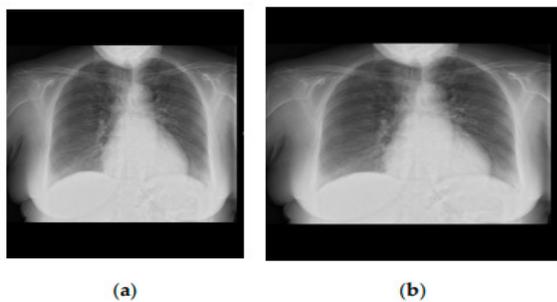


Fig. 3. Distortion Due to the Images being Resized to a Square Shape. (a) Displays the Original CXr Image of a COVID Patient; (b) Displays the Resampled Image in a Square Shape.

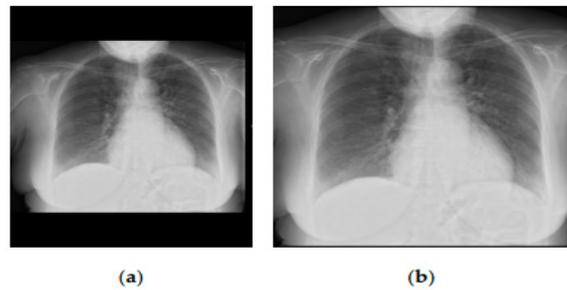


Fig. 4. Pre-Processing is Applied to All Dataset Files. (a) Displays the Original CXr Image of an Individual with COVID-19 (b) Displays a Square Shape of the Pre-Processed Image.

is:

$$F(i, j) = (J \times K)(i, j) = \sum \sum J(i + p, j + q)K(p, q), \quad (1)$$

where the input matrix is represented with J , a 2D filter of size $p \times q$ is indicated by K , and F denotes a 2D feature map's output. $J \times K$ describes the operation of the convolutional layer.

A.2 Rectified Linear Unit: Each convolution layer has an activation function. Here this model used the ReLU activation function. To extend nonlinearity in feature maps, the ReLU layer is utilized [33]. ReLU itemizes activation by maintaining the threshold input at zero. It is mathematically represented as follows:

$$f(y) = \max(0, y) \quad (2)$$

A.3 Zero Padding Layer: The application of this layer is to add columns and rows of zeros to the left, right, top, and

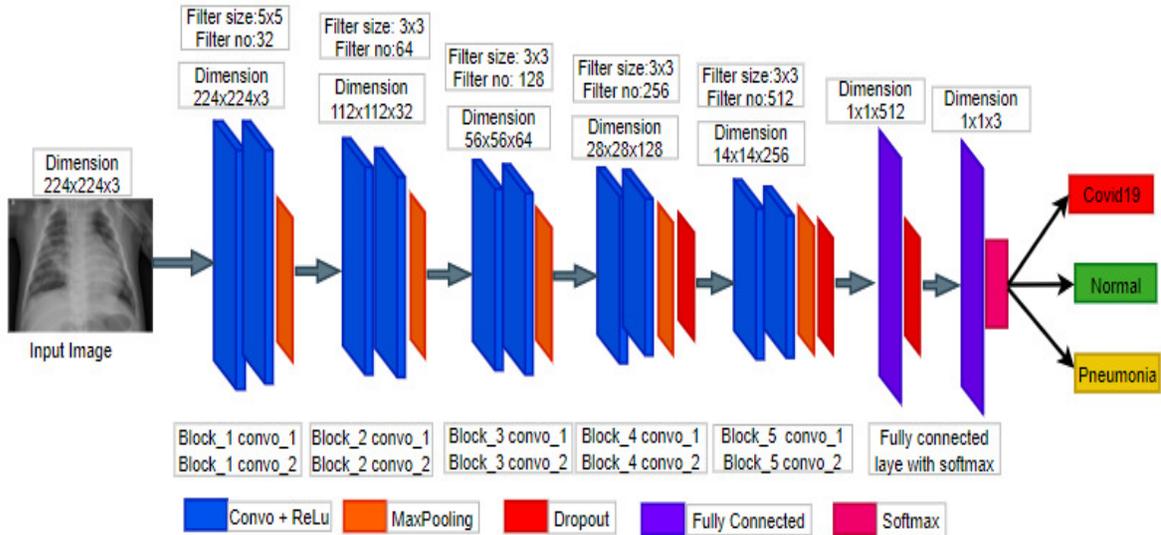


Fig. 5. Proposed CNN Architecture.

bottom sides of an image. We have applied 1×1 zero padding during our work.

A.4 Pooling Layers : The next layer that we applied is a pooling layer. This layer is used to reduce the number of consumption and parameters in the network by dynamically shrinking the spatial size of the images. The pooling layer aids in reducing the problem of overfitting and executes distinct operations on each depth slice of the input and resizes it spatially using the MAX function. The most common MAX pooling layer with filters of size 2×2 is used with a stride of 2. MAX pooling is applied to downsample the input by two along width and height, discarding 75% of the activation. The pooling layer is represented below:

- 1) Accepted a volume of size is $w1 \times h1 \times d1$
- 2) Required two hyper parameters:
 - F is their spatial extent
 - S is the stride
- 3) Produced a volume of size is $w2 \times h2 \times d2$:
 - $w2 = (w1 \times F) / S + 1$
 - $h2 = (h1 \times F) / S + 1$
 - $d2 = d1$
- 4) It presents zero parameters since it calculates a specified function of the input.
- 5) Zero padding is not commonly used in the pooling layers [34].

A.5 Dropout Layer : There are many methods to handle a CNN's capacity to eliminate overfitting. For example, the dropout layer [35] is an effective regularization method to prevent overfitting problems. During training time, the dropout layer randomly sets some neuron's activation to zero and those neurons will not update their weights. Dropout is activated with some probability during training, otherwise it is inactive. Some

neurons do not learn all characteristics due to dropout. There is no dropout layer added throughout the testing period. Most of the times, a dropout ratio $p = 0.5$ shows good results, hence is selected as default. The p value can be tuned during data validation as well.

A.6 Fully Connected Layer: Each neuron in one layer is connected to another neuron in another layer, hence FCLs are formed. The layer basically takes an input image or object and outputs an N -dimensional vector as a response, where N is the number of given classes that the model or program has to choose from [36]. The working technique is as follows: the system looks at the output of the preceding layer and identifies which features are most commonly associated with a specific class. Fundamentally, a FCL considers which high-level features are most closely associated with a certain class and assigns weights to them, resulting in the classification value when the products of the weights and the preceding layer output are computed.

A.7 Softmax : The softmax function with loss [37] is used, which crushes an N -dimensional vector x of random real values to an N -dimensional vector ($\sigma(x)$) of real values in the range from 0 to 1 that will sum up to 1. The function is given by the following equation 3:

$$\sigma(x)_j = \frac{e^z_j}{\sum_{k=1}^k e^z_j} \quad (3)$$

Here, $j = 1, 2, \dots, k$. For this model, the output feature map will be for three classes. In the output feature map, the pixels belonging to the predicted class will be 1, and for that same pixel, other classes will contain zero.

4) *Cost-Sensitive Learning*: Our proposed method focuses on learning features automatically from CXR images based on CNN. Cost-sensitive learning (CSL) [38] is a mechanism in which the method that ranks each class in a classification problem can be penalized. When using CNNs, CSL helps prevent prejudices in classification and then assists in overcoming the issue of imbalance as a tool. We have chosen CNN as our data set is imbalanced. The approach of class weight is used in this research as one possible way to mitigate the future effects of data imbalance. We adjust the weights in the class weight system in inverse proportion to the number of occurrences of each class in the raw data. We have used a Sci-Kit Learn feature for all this, which acquires numbers to match the number of cases centred on logistic regression concepts [39][40]. The weight W_k in class k is calculated using the following equation 4.

$$W_k = \frac{\sum \text{Cases}}{\sum \text{Classes} \times \sum \text{Cases in class}(k)} \quad (4)$$

The weights of the groups are utilized to match the standard. As a result, in the loss function, we assign higher values to cases involving smaller groups. The estimated loss would then be a weighted average, with W_k denoting the weights assigned to each class inside each sample throughout the loss calculation.

5) *Loss Function Used: Categorical Cross-Entropy*: The objective of network training is to maximize the probability of correct classification. This is gained by minimizing the cross-entropy loss for each training sample. The loss function utilized in our work is the categorical cross-entropy loss. Our cases are divided into three categories. The failure is determined independently for each class, and the values are then combined together. In the equation 5 [41] below, categorical cross-entropy is specified.

$$L(p, q) = - \sum W_k \times p_k \times \log(q_k), \quad (5)$$

where k is the class number (in this work, the classes are COVID-19, Normal, and Pneumonia), q is the predicted probability or softmax function of class k and W_k is the weight for class k .

6) *Screening and Localizing Pathogens*: In order to better understand the stimulation of the last convolutionary layer of the developed model, we have applied the Gradient-Weighted Classes Activation Mapping (Grad-CAM) [42] algorithm. This layer is the only one that delivers the parameters for the logistic layer of the final parameters to determine a probability distribution output. For constructing the Grad-CAM of an input file, the gradients of this layer have been used. Grad-CAM thus gives a coarse localization map that shows the most significant areas in the picture (radiological features).

IV. EXPERIMENTAL RESULTS AND DISCUSSION

In this section, we discussed the implementation phase and evaluated the classification efficiency of the proposed architecture. The performances and the characteristics of our proposed system are also compared with existing works for COVID-19 detection.

A. Experimental Process

The experiment performed in this research are described in this section.

1) *Experimental Environment*: The proposed approach uses a 0.001 learning rate and a 25-epoch number. On the 1.80 GHz Intel (R) Core i5-8265U Processor, the proposed CNN was developed utilizing *Keras* and the *Python* package with the *TensorFlow2* backend. Additionally, the experiments were conducted utilizing Google Colab's graphical processing unit (GPU).

2) *Split Data into Training set and Testing set*: We divide the data into training data, consisting of 90% (9264 images) and testing that makes 10% (1029 images). For validation, we further divide the training data again as 10%. The images were distributed randomly across the test and training datasets, ensuring that the two datasets are shared roughly evenly between the two groups. It is also crucial that the model is trained in a variety of ways. To put it in another way, it would be overfitted, which is not a good thing.

3) *Model Training Methodology*: The Adam optimizer was used to train every layer of the improvement parcel with the normal performance parameters ($\beta_1 = 0.9$ and $\beta_2 = 0.999$) using batches of size 32 and running throughout every epoch through the whole dataset. Eventually, adding an effective learning rate scheduler called "ReduceLROnPlateau" defines a lookup to control the relevance of the loss of reliability. It starts with a learning rate of 0.001 on demand.

We reduced the hyperparameter learning rate by a factor of 0.5 after 5 epochs, with no improvement recorded by the detector. We have also included an early stop to prevent the network from over-learning in the regularization step. We stopped training the model after 5 epochs because the loss score had not changed. The hyperparameters used in the proposed model are shown in Table IV.

B. Performance Evaluation Metrics

Five metrics, i.e. accuracy, sensitivity, specificity, F1-score, and precision, are used to assess the proposed system's performance. Some variables are employed in the metrics calculation, as mentioned below:

- True Positives (TPs) are the COVID-19 cases correctly estimated;
- False Positives (FPs) are pneumonia or normal cases that are misclassified as COVID-19;
- True Negatives (TNs) are correctly classified pneumonia or normal cases;
- False Negatives (FNs) are COVID-19 cases that are misclassified as pneumonia or normal cases.

Using the above mentioned variables, performance evaluation metrics can be computed as:

- Accuracy: A metric that normally expresses how the model performs across all classes. It is determined as the proportion between the quantity of right prediction to the all the number of predictions, i.e.:

TABLE IV. HYPER PARAMETERS OF THE NETWORK

Hyper Parameter	Weight
Batch size	32
Cost function	Categorical Cross Entropy
Learning Rate (LR)	0.001
LR Multiplying factor	0.5
LR Decay	5 times after a plateau
Epochs	25
Optimizer	Adam

$$Accuracy = \frac{(TN + TP)}{(TP + FP + TN + FN)} \quad (6)$$

- Sensitivity: It indicates the ratio of individuals who test positive among all those who have the diseases, expressed as:

$$Sensitivity = \frac{TP}{(TP + FN)} \quad (7)$$

- Specificity: The ratio of individuals who test negative among all those that actually do not have that disease, expressed as:

$$Specificity = \frac{TN}{(TN + FP)} \quad (8)$$

- F1-score: The F1-score is a count of a test's accuracy. This score may be interpreted as a weighted average of the precision and defined as:

$$F1 - Score = \frac{2TP}{(2TP + FN + FP)} \quad (9)$$

- Precision: The proportion of the correctly predicted positive individuals to the total number of actual positive individuals and calculated as:

$$Precision = \frac{TP}{(TP + FP)} \quad (10)$$

It is required to evaluate classification findings using graphical approaches such as the ROC curve and its general ranking, and the region below, i.e., the AUC.

C. Results Analysis

The confusion matrix enables us to quantify the metrics of our categorization study's results. The confusion matrix for the test cases of the proposed CNN framework is illustrated in Fig. 6.

The suggested approach misclassified 44 of the 1029 test images, with adequate and consistent true negative and true positive scores. As a result, the suggested CNN architecture can effectively classify COVID-19 cases.

Furthermore, the CNN classifier performance evaluation is graphically represented in Fig. 7 in terms of loss and accuracy within the validation and training phases. At epoch number 25, the training, as well as validation accuracy, is 96.6% and 95.3%, respectively. Furthermore, the validation and training loss achieved by the proposed system are 0.205 and 0.102, respectively.

The general accuracy, sensitivity, specificity, F1-score, and precision for every issue of the proposed system are

Predict	PNEUMONIA	82	1	6
	COVID19	1	232	19
	NORMAL	0	17	671
		PNEUMONIA	COVID19	NORMAL
		Actual		

Fig. 6. Confusion Matrix for the Proposed Architecture.

summarized in Table V. The proposed CNN has accuracy 96.3%, sensitivity 93%, specificity 97.4%, F1-score 92%, and precision 92% for COVID-19 cases. For normal cases, accuracy 99.2%, sensitivity 99%, specificity 99.3%, F1-score 92%, and precision 95% have been recorded. The pneumonia cases have obtained an accuracy of 95.9%, sensitivity 96%, specificity 94.9%, F1-score 97%, and precision 98%. The highest accuracy, specificity, and sensitivity are obtained within the normal cases. For example, the best F1-Score and precision have been found in pneumonia cases, along with the highest specificity value.

For better understanding, the classification measurements for every class in terms of accuracy, specificity, precision, recall, and F1-score are shown visually in Fig. 8.

In addition, ROC curves are presented between the false positive rate and thus the true positive rate to check the general performance, which is shown in Fig. 9. The AUC has been determined to be 96.6% for the proposed CNN architecture.

Experimental findings show that the proposed architecture has achieved 96.3% accuracy, 96% AUC, 97.4% specificity, 92% F1-score, 92% precision, and 93% sensitivity for the COVID-19 infected cases.

1) *Screening for Pathogens:* Grad-CAM is often referred to as a heat map that uses the gradients of an identifying and mapping to view our test dynamically. In order to highlight the significant portions within the image for forecasting, a rough localization map is generated after passing into the last layer. The most important region (processed attributes) from which

TABLE V. PERFORMANCE OF THE PROPOSED NETWORK (IN %)

Class	Accuracy	Sensitivity	Specificity	Precision	F1-Score
Normal	99.2	99	99.3	92	95
Pneumonia	95.9	96	94.9	98	97
Covid19	96.3	93	97.4	92	92

TABLE VI. COMPARATIVE ANALYSIS OF THE PROPOSED ARCHITECTURE WITH EXISTING ONES IN TERMS OF THE SELECTED PERFORMANCE METRICS (IN %)

Author	Accuracy	Specificity	Sensitivity	F1-Score	Precision
Alqudah et al. [14]	95.2	100	93.3	92	90
Kumar et al. [15]	95	96	95	95	95
Apostolopoulos et al. [18]	94.72	96.46	98.66	91	91.5
Khan et al. [20]	95	97.5	96.9	95.6	95
Hemdan et al. [21]	90	89	90	90	83
Sing et al. [22]	96	92	95	96	91
Li et al. [23]	90	93	90	90	91
Proposed System	96.3	97	96	96	96

TABLE VII. COMPARISON OF THE PROPOSED METHOD WITH EXISTING WORKS REGARDING DETECTION ACCURACY

Authors	Classes	Dataset	Accuracy(%)
Alqudah et al. [14]	2-classes	23 NonCOVID-19/48 COVID-19	95.2%
Apostolopoulos et al. [18]	3-classes	700 Pneumonia/504 Normal/224 COVID-19	93.48%
Hemdan et al. [21]	2-classes	25 non-COVID-19/25 COVID-19	90%
R Kumar et al. [15]	3-classes	1345 Pneumonia/1341 Normal/62 COVID-19	90%
Sethy and Behera [43]	3-classes	127 Pneumonia/127 Normal/127 COVID-19	95.33%
Panwar et al. [44]	2-classes	142 Normal/142 COVID-19	88%
Wang and Wong [11]	3-classes	8066 Pneumonia/5538 Normal/358 COVID-19	93.3%
Ozturk et al. [45]	3-classes	500 Pneumonia/500 Normal/125 COVID-19	87.02%
Proposed System	3-classes	4200 Pneumonia/3218 Normal/2875 COVID-19	96.3%

the classification determination has been made by the network are shown in a deep blue colour. Fig. 10 shows the heat map for normal, COVID-19, and pneumonia cases of classified test samples.

D. Comparative Assessment

The investigation of the outcomes demonstrates that a CNN architecture is highly effective in detecting COVID-19 by sustaining automated feature removal from CXR images. As a result, our proposed approach can separate COVID-19 from normal and pneumonia cases with significant accuracy.

Moreover, to compare the presented framework with existing frameworks, we have re-constructed the methods used in [15], [20], [22] and [23]. All the models are evaluated on the same training, validation, and testing data set to guarantee the fairness of the comparison. Table VI shows a comparative analysis of our system with the above works. Graphical comparisons of the proposed approach with [15], [20], [22] and [23] is shown in Fig. 11.

To assess the performance of our suggested system, we compared the execution results of some existing models to the results of our model, as shown in Table V.

As shown in Table VII, our suggested COVID-19 detection model outperformed a number of other state-of-the-art detection frameworks. Note that it is not feasible to check the performance of the suggested framework with other existing frameworks because of the number of test images; furthermore, because the data sources do not seem to be the same. Accordingly, the correlation between these frameworks is for expository purposes. However, it demonstrates the capability of our methodology in this undertaking. From Table VII, we can

state that the results of our proposed framework are generally better than most of the existing frameworks. More specifically, our framework achieves the highest accuracy, F1-score and precision, while it is within a percentage point of the maximum specificity and sensitivity achieved by other methods.

V. CONCLUSION

This study designed a deep CNN-based system to classify chest X-ray images to detect COVID-19 cases, pneumonia, and healthy cases. To identify COVID-19 cases, we used CNN as a motif identifier. The proposed approach has successfully isolated COVID-19 cases from the normal cases and can produce 97% specificity, 96.3% accuracy, 96% sensitivity, 96% F1-score, and 96% precision. The empirical result shows that our proposed framework achieves better accuracy than existing works, and the radiologists assess these performances. In addition, the system is ready to be tested with a more extensive database. It can be employed in local areas where people are affected by the COVID-19 to overcome the lack of expert radiologists.

The suggested framework has a limitation in using a small number of X-ray images for the COVID-19 cases. Future work, which is already under way at the time of writing, includes running our experiments on more such X-ray images from remote hospitals to make our model robust and effective.

ACKNOWLEDGMENT

This work was partially supported by the Italian Ministry of Education and Research (MIUR) in the framework of the CrossLab project (Departments of Excellence).

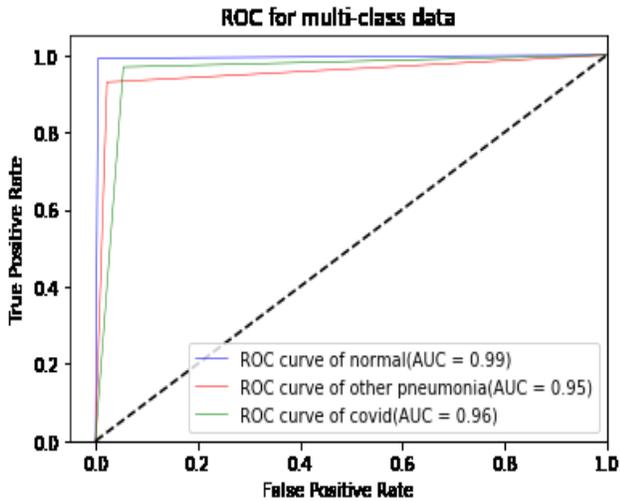


Fig. 9. ROC Analysis of the Proposed Architecture.

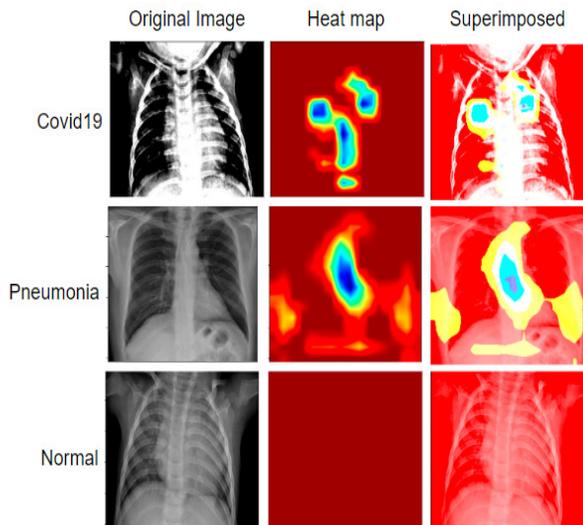


Fig. 10. Heat Map Image Created by the Proposed Architecture.

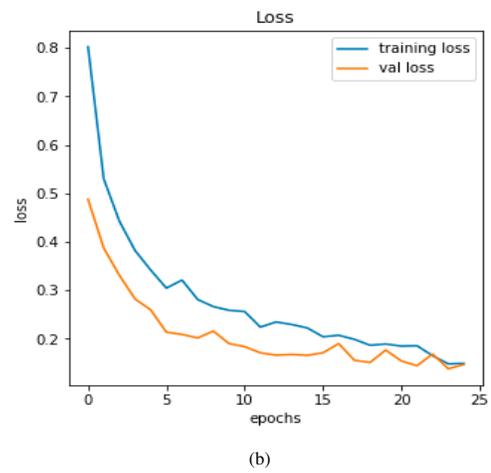
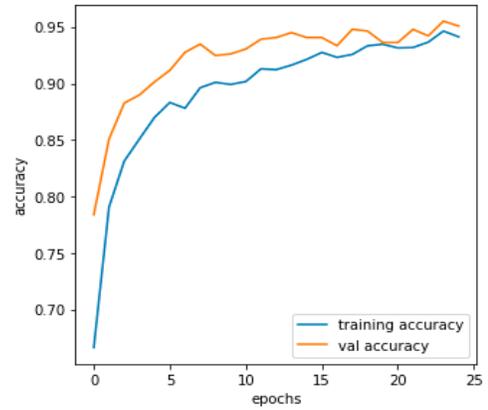


Fig. 7. Performance Evaluation Metrics of the Suggested Approach (a) Accuracy Graph, (b) Loss Graph.

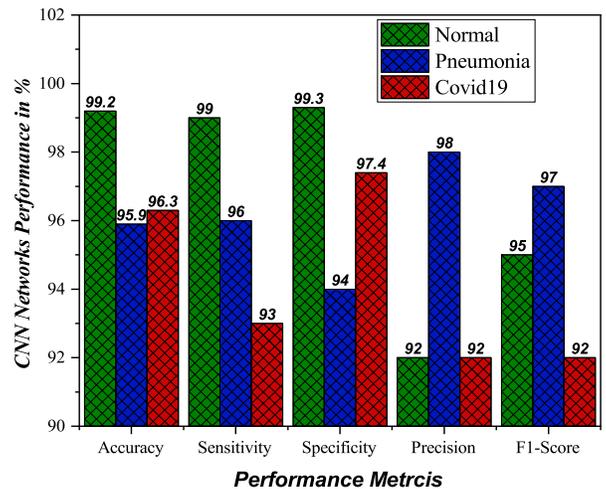


Fig. 8. Performance of the Proposed CNN.

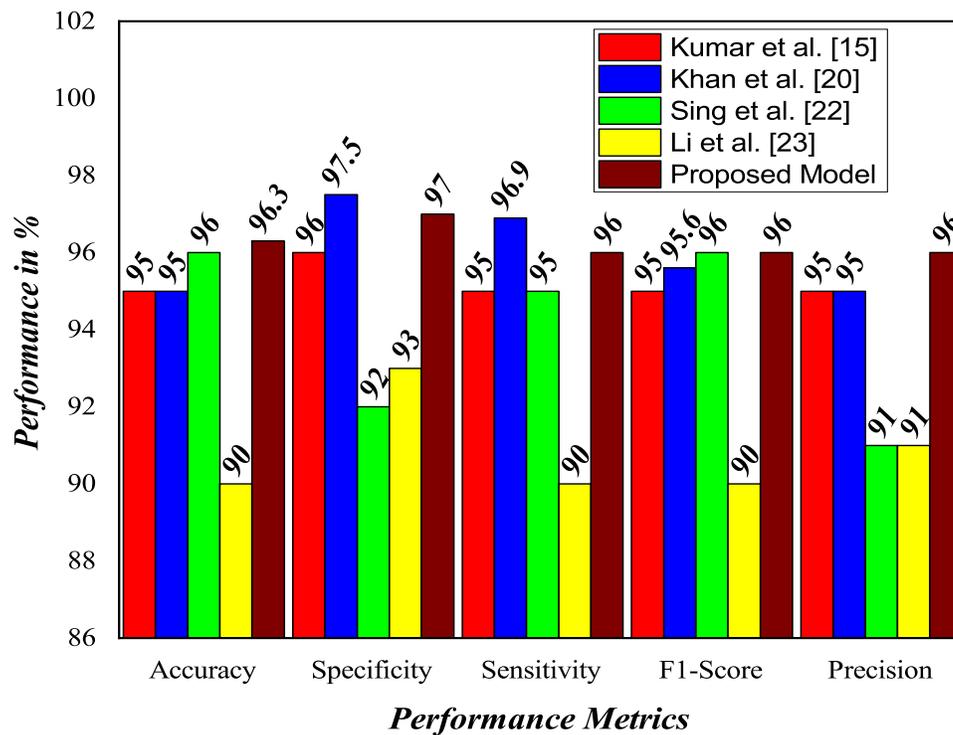


Fig. 11. Graphical Comparison of the Proposed Model with [15], [20] and [23] when Trained and Tested on our Datasets.

REFERENCES

- [1] T. Singhal, "A review of coronavirus disease-2019 (covid-19)," *The Indian Journal of Pediatrics*, pp. 1–6, 2020.
- [2] C.-C. Lai, T.-P. Shih, W.-C. Ko, H.-J. Tang, and P.-R. Hsueh, "Severe acute respiratory syndrome coronavirus 2 (sars-cov-2) and corona virus disease-2019 (covid-19): the epidemic and the challenges," *International journal of antimicrobial agents*, p. 105924, 2020.
- [3] J. Li, J. J. Li, X. Xie, X. Cai, J. Huang, X. Tian, and H. Zhu, "Game consumption and the 2019 novel coronavirus," *The Lancet Infectious Diseases*, vol. 20, no. 3, pp. 275–276, 2020.
- [4] J. M. Sharfstein, S. J. Becker, and M. M. Mello, "Diagnostic testing for the novel coronavirus," *Jama*, vol. 323, no. 15, pp. 1437–1438, 2020.
- [5] F. A. Rabi, M. S. Al Zoubi, G. A. Kasasbeh, D. M. Salameh, and A. D. Al-Nasser, "Sars-cov-2 and coronavirus disease 2019: what we know so far," *Pathogens*, vol. 9, no. 3, p. 231, 2020.
- [6] A. York, "Novel coronavirus takes flight from bats?" *Nature Reviews Microbiology*, vol. 18, no. 4, pp. 191–191, 2020.
- [7] M. Loey, F. Smarandache, and N. E. M. Khalifa, "Within the lack of chest covid-19 x-ray dataset: A novel detection model based on gan and deep transfer learning," *Symmetry*, vol. 12, no. 4, p. 651, 2020.
- [8] V. M. Corman, O. Landt, M. Kaiser, R. Molenkamp, A. Meijer, D. K. Chu, T. Bleicker, S. Brünink, J. Schneider, M. L. Schmidt *et al.*, "Detection of 2019 novel coronavirus (2019-ncov) by real-time rt-pcr," *Eurosurveillance*, vol. 25, no. 3, p. 2000045, 2020.
- [9] W.-j. W. Liu, C. Yuan, M.-l. Yu, P. Li, and J.-b. Yan, "Detection of novel coronavirus by rt-pcr in stool specimen from asymptomatic child, china," *Emerg Infect Dis J*, 2020.
- [10] W. Wang, Y. Xu, R. Gao, R. Lu, K. Han, G. Wu, and W. Tan, "Detection of sars-cov-2 in different types of clinical specimens," *Jama*, vol. 323, no. 18, pp. 1843–1844, 2020.
- [11] L. Wang, Z. Q. Lin, and A. Wong, "Covid-net: A tailored deep convolutional neural network design for detection of covid-19 cases from chest x-ray images," *Scientific Reports*, vol. 10, no. 1, pp. 1–12, 2020.
- [12] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1026–1034.
- [13] M. Rahimzadeh and A. Attar, "A modified deep convolutional neural network for detecting covid-19 and pneumonia from chest x-ray images based on the concatenation of xception and resnet50v2," *Informatics in medicine unlocked*, vol. 19, p. 100360, 2020.
- [14] A. M. Alqudah, S. Qazan, H. Alquran, I. A. Qasmieh, and A. Alqudah, "Covid-2019 detection using x-ray images and artificial intelligence hybrid systems," *Biomedical Signal and Image Analysis and Project; Biomedical Signal and Image Analysis and Machine Learning Lab: Boca Raton, FL, USA*, 2019.
- [15] R. Kumar, R. Arora, V. Bansal, V. J. Sahayashela, H. Buckchash, J. Imran, N. Narayanan, G. N. Pandian, and B. Raman, "Accurate prediction of covid-19 using chest x-ray images through deep feature learning model with smote and machine learning classifiers," *medRxiv*, 2020.
- [16] M. J. Horry, M. Paul, A. Ulhaq, B. Pradhan, M. Saha, N. Shukla *et al.*, "X-ray image based covid-19 detection using pre-trained deep learning models," *engrXiv*, 2020.
- [17] F. Ucar and D. Korkmaz, "Covidagnosis-net: Deep bayes-squeezenet based diagnostic of the coronavirus disease 2019 (covid-19) from x-ray images," *Medical Hypotheses*, p. 109761, 2020.
- [18] I. D. Apostolopoulos and T. A. Mpesiana, "Covid-19: automatic detection from x-ray images utilizing transfer learning with convolutional neural networks," *Physical and Engineering Sciences in Medicine*, p. 1, 2020.
- [19] S. K. Bandyopadhyay and S. Dutta, "Machine learning approach for confirmation of covid-19 cases: Positive, negative, death and release," *medRxiv*, 2020.
- [20] A. I. Khan, J. L. Shah, and M. M. Bhat, "Coronet: A deep neural network for detection and diagnosis of covid-19 from chest x-ray images," *Computer Methods and Programs in Biomedicine*, p. 105581, 2020.
- [21] E. El-Din Hemdan, M. A. Shouman, and M. E. Karar, "Covidx-net: A framework of deep learning classifiers to diagnose covid-19 in x-ray

- images," *arXiv*, pp. arXiv-2003, 2020.
- [22] M. Singh, S. Bansal, S. Ahuja, R. K. Dubey, B. K. Panigrahi, and N. Dey, "Transfer learning based ensemble support vector machine model for automated covid-19 detection using lung computerized tomography scan data," *Research Square*, 2020.
- [23] X. Li, C. Li, and D. Zhu, "Covid-mobilexpert: On-device covid-19 screening using snapshots of chest x-ray, 2020," 2020.
- [24] S. Ahuja, B. K. Panigrahi, N. Dey, V. Rajinikanth, and T. K. Gandhi, "Deep transfer learning-based automated detection of covid-19 from lung ct scan slices," *Applied Intelligence*, vol. 51, no. 1, pp. 571–585, 2021.
- [25] S. J. Fong, G. Li, N. Dey, R. G. Crespo, and E. Herrera-Viedma, "Composite monte carlo decision making under high uncertainty of novel coronavirus epidemic using hybridized deep learning and fuzzy rule induction," *Applied Soft Computing*, p. 106282, 2020.
- [26] P. Afshar, S. Heidarian, F. Naderkhani, A. Oikonomou, K. N. Plataniotis, and A. Mohammadi, "Covid-caps: A capsule network-based framework for identification of covid-19 cases from x-ray images," *Pattern Recognition Letters*, vol. 138, pp. 638–643, 2020.
- [27] H. Panwar, P. Gupta, M. K. Siddiqui, R. Morales-Menendez, P. Bhardwaj, and V. Singh, "A deep learning and grad-cam based color visualization approach for fast detection of covid-19 cases using chest x-ray and ct-scan images," *Chaos, Solitons & Fractals*, vol. 140, p. 110190, 2020.
- [28] R. Karthik, R. Menaka, and M. Hariharan, "Learning distinctive filters for covid-19 detection from chest x-ray using shuffled residual cnn," *Applied Soft Computing*, vol. 99, p. 106744, 2021.
- [29] J. P. Cohen, P. Morrison, and L. Dao, "Covid-19 image data collection," *arXiv 2003.11597*, 2020. [Online]. Available: <https://github.com/ieee8023/covid-chest-xray-dataset>
- [30] P. Patel, "Chest x-ray images (pneumonia)," 2020. [Online]. Available: <https://www.kaggle.com/prashant268/chest-x-ray-covid19-pneumonia>
- [31] P. Mooney, "Chest x-ray images (pneumonia)," 2017. [Online]. Available: <https://www.kaggle.com/paultimothy-mooney/chest-x-ray-pneumonia>
- [32] F. Pasa, V. Golkov, F. Pfeiffer, D. Cremers, and D. Pfeiffer, "Efficient deep network architectures for fast chest x-ray tuberculosis screening and visualization," *Scientific reports*, vol. 9, no. 1, pp. 1–9, 2019.
- [33] D. Singh, V. Kumar, and M. Kaur, "Classification of covid-19 patients from chest ct images using multi-objective differential evolution-based convolutional neural networks," *European Journal of Clinical Microbiology & Infectious Diseases*, pp. 1–11, 2020.
- [34] H. Kutlu and E. Avci, "A novel method for classifying liver and brain tumors using convolutional neural networks, discrete wavelet transform and long short-term memory networks," *Sensors*, vol. 19, no. 9, p. 1992, 2019.
- [35] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *The journal of machine learning research*, vol. 15, no. 1, pp. 1929–1958, 2014.
- [36] P. Chang, J. Grinband, B. Weinberg, M. Bardis, M. Khy, G. Cadena, M.-Y. Su, S. Cha, C. Filippi, D. Bota *et al.*, "Deep-learning convolutional neural networks accurately classify genetic mutations in gliomas," *American Journal of Neuroradiology*, vol. 39, no. 7, pp. 1201–1207, 2018.
- [37] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Semantic image segmentation with deep convolutional nets and fully connected crfs," *arXiv preprint arXiv:1412.7062*, 2014.
- [38] V. López, A. Fernández, J. G. Moreno-Torres, and F. Herrera, "Analysis of preprocessing vs. cost-sensitive learning for imbalanced classification. open problems on intrinsic data characteristics," *Expert Systems with Applications*, vol. 39, no. 7, pp. 6585–6608, 2012.
- [39] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg *et al.*, "Scikit-learn: Machine learning in python," *the Journal of machine Learning research*, vol. 12, pp. 2825–2830, 2011.
- [40] G. King and L. Zeng, "Logistic regression in rare events data," *Political analysis*, vol. 9, no. 2, pp. 137–163, 2001.
- [41] K. P. Murphy, *Machine learning: a probabilistic perspective*. MIT press, 2012.
- [42] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 618–626.
- [43] P. K. Sethy and S. K. Behera, "Detection of coronavirus disease (covid-19) based on deep features, 2020," 2020.
- [44] H. Panwar, P. Gupta, M. K. Siddiqui, R. Morales-Menendez, and V. Singh, "Application of deep learning for fast detection of covid-19 in x-rays using ncovnet," *Chaos, Solitons & Fractals*, p. 109944, 2020.
- [45] T. Ozturk, M. Talo, E. A. Yildirim, U. B. Baloglu, O. Yildirim, and U. R. Acharya, "Automated detection of covid-19 cases using deep neural networks with x-ray images," *Computers in Biology and Medicine*, p. 103792, 2020.

BERT-based Approach to Arabic Hate Speech and Offensive Language Detection in Twitter: Exploiting Emojis and Sentiment Analysis

Maha Jarallah Althobaiti
Department of Computer Science
College of Computers and Information Technology
Taif University
Taif 21944, Saudi Arabia

Abstract—The user-generated content on the internet including that on social media may contain offensive language and hate speech which negatively affect the mental health of the whole internet society and may lead to hate crimes. Intelligent models for automatic detection of offensive language and hate speech have attracted significant attention recently. In this paper, we propose an automatic method for detecting offensive language and fine-grained hate speech from Arabic tweets. We compare between BERT and two conventional machine learning techniques (SVM, logistic regression). We also investigate the use of sentiment analysis and emojis descriptions as appending features along with the textual content of the tweets. The experiments shows that BERT-based model gives the best results, surpassing the best benchmark systems in the literature, on all three tasks: (a) offensive language detection with 84.3% F1-score, (b) hate speech detection with 81.8% F1-score, and (c) fine-grained hate-speech recognition (e.g., race, religion, social class, etc.) with 45.1% F1-score. The use of sentiment analysis slightly improves the performance of the models when detecting offensive language and hate speech but has no positive effect on the performance of the models when recognising the type of the hate speech. The use of textual emoji description as features can improve or deteriorate the performance of the models depending on the size of the examples per class and whether the emojis are considered among distinctive features between classes or not.

Keywords—Deep learning, hate speech detection; offensive language detection; sentiment analysis; transformer-based model; BERT; emoji

I. INTRODUCTION

The pervasiveness of hatred and offensive content on the internet has become disturbing, raising an alarm over negative consequences for the target individuals' mental health and the internet society's well-being [1], [2]. Online hateful and offensive language detection aims to make the internet not only accessible but also safe, as hateful speech online threatens society by encouraging hate crimes [3]. It also enables the scientific analyses of such abusive languages, covers their causes, and establishes possible solutions. Thus, in recent years, Artificial Intelligence (AI) and Natural Language Processing (NLP) communities have investigated various techniques as potential solutions for automatically detecting offensive language online with high performance [4], [5], [6], [7], [8], [9], [10].

Recently, a series of workshops and shared tasks have been conducted to explore the problem from various perspectives.

Significant attention has been given to defining the problem and investigating the automatic detection techniques of offensive language with all its types and ways, including abuse, aggression, cyberbullying, and hateful content. For example, in 2018, there was the first workshop on Trolling, Aggression and Cyberbullying (TRAC-2018) [11], [12]. In addition, there have been a series of five workshops on online abusive language and harms since 2017 [13], [14], [15], [16], [17]. The sixth edition of this workshop (6th WOA) will be held on July 14th with the 2022 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL) [18]. Two editions of shared tasks on offensive language identification were organised at the international workshop on Semantic Evaluation (SemEval) in 2019 [19] and 2020 [20]. Regarding Arabic, there was a shared task on offensive language detection for Arabic at the 4th workshop on Open-Source Arabic Corpora and Processing Tools (OSACT4) [21]. Another shared task on fine-grained hate speech detection on Arabic Twitter will be held on 20 June at the 5th workshop on Open-Source Arabic Corpora and Processing Tools (OSACT5) co-located with LREC 2022 [22].

Several categories have been adopted to define aggressive languages in online content. Among them, [23] classifies online content into hate speech, offensive, neither offensive nor hate-speech, while [24] classifies online content into abusive, hateful, normal, or spam. The study of [25] classifies online comments as racist, sexist, or neither. In addition, [4] proposed a typology of all works that have been grouped under the label of hate speech, cyberbullying, and online abuse. They synthesised the work on online abusive language in a two-fold typology that considers whether (a) the abuse is directed at a specific target and (b) the degree to which it is explicit.

In this study, we propose methods for the following three tasks: (a) offensive language detection (identifying whether a tweet is offensive or not), (b) hate speech detection (i.e., identifying whether a tweet has hate speech or not), and (c) fine-grained hate speech detection (i.e., identifying and recognising the type of hate speech: disability, gender, ideology, race, religion, or social class). We utilise the dataset released by [26] which contains 12,698 Arabic tweets annotated for the three aforementioned tasks. We also investigate the use of two conventional machine learning techniques: Support Vector Machine (SVM) and Logistic Regression (logit). We also explore Bidirectional Encoder Representations from Trans-

formers (BERT), a state-of-the-art transformer-based machine learning technique for deep-contextualised word representation. In addition, sentiment analysis and emoji description are explored as potential features that can be utilised in training any model to improve its performance, as our intuition indicates that hate speech and offensive language mostly express negativity which can be exploited. The contributions of this study are summarised as follows.

- We investigate and compare conventional machine-learning techniques and a transfer-based model (BERT) for offensive language and fine-grained hate speech detection.
- We examine the use of sentiment analysis and emoji descriptions as additional textual features for both transformer-based models and conventional machine learning methods.
- We examine our proposed methods on relatively small unbalanced data and with different preprocessing settings.
- We develop a novel and simple method for offensive language and hate-speech detection that outperforms the best benchmark systems reported in the literature with the released dataset used in our study.

II. RELATED WORK

A considerable number of studies for detecting online hate speech and offensive language have been suggested and investigated in the literature in the past ten years but intensively since 2017 [27], [28], [29], [30], [7], [31]. Workshops and shared tasks organised for the task of detecting and recognising online offensive language, hate-speech, and abusive content played a vital role in attracting the attention of the research community to propose potential techniques for the task [14], [13], [11], [15], [20], [18]. Many studies have examined generalised solutions for offensive language detection from online content in multiple languages, while other studies concentrated on examining the suitable features and techniques for one language, such as Greek [32], Chinese [33], Slovene [34], and Croatian [35]. Significant attention has been paid to detect offensive language from online English content [36], [5].

Few studies have been conducted to address the problem of online anti-social behaviour on Arabic; most have targeted offensive language detection, while the remaining studies have investigated the problem of hate speech detection [21], [37], [38], [39]. One of the early works, conducted by [38], targeted vulgar and pornographic obscene speech on Arabic social media using a list-based approach. They used tweets to build a list of seed words for obscene phrases. Then, they employed the list to construct three sublists of obscene words and phrases using multiple measurements, such as the Log Odds Ratio (LOR) for unigrams and bigrams. Conventional Machine Learning (ML) techniques have also been investigated for offensive language and hate speech detection [40], [41]. The most commonly used traditional ML techniques for offensive language and hate-speech detection are SVM [37], [39], [42], [43], [44], Naive Bayes [43], [44], and Logistic Regression (logit) [45], [46].

The study in [47] examined the use of FastText Deep Learning (DL) model on a dataset containing 36 million tweets to detect offensive speech. They reported that the FastText DL model outperformed an SMV classifier trained on character n-gram features. Mohaouchane, Mourhir, and Nikolov [48] explored the use of AraVec word embeddings and four DL models: Bidirectional Long Short-Term Memory (Bi-LSTM), Bi-LSTM with an attention mechanism, Convolutional Neural Network (CNN), and a combined model of CNN and LSTM. The experiments illustrated the outperforming results of the CNN over all other models. Many other architectures of deep neural networks have been investigated, such as Gated Recurrent Unit (GRU) [49], RNN [41], [45], and contextual embeddings (e.g., multilingual BERT [46], [50] and AraBERT [51]).

The results of the techniques in most of the aforementioned studies cannot be compared because every study used their own dataset and the available datasets for Arabic offensive language and hate-speech detection are limited [39], [43], [44]. However, the shared task on offensive language and hate speech detection in the fourth workshop on Open-Source Arabic Corpora and Corpora Processing Tools (OSACT4) provides a manually annotated Twitter dataset, consisting of 10,000 examples, for offensiveness (labels are: OFF or NOT_OFF) and for specifying the offensive content type of an offensive example as hate speech or not (labels are: HS or NOT_HS). This provides an opportunity to compare techniques for both tasks: offensive language detection and hate speech detection. The winning team for Arabic offensive language detection has employed an ensemble system of traditional machine learning technique (SVM) and two DL models: CNN+BiLSTM and multilingual BERT with an F1-score equal to 90.51%, while the best performing system for hate speech detection used SVM and achieved 95.2%, outperforming the second-place system by 12.9%. The winning team has attributed the performance of the winning model to the intensive preprocessing steps which included emoticons and emoji to textual description conversion, dialectal to MSA conversion, word categorisation (e.g., all animal names included in tweets were reduced to only one word.), letter normalisation, stop-word removal, and hashtag segmentation.

A study by Mubarak et al. [26] released an Arabic dataset for detecting offensive language and hate speech, consisting of 12,698 tweets. To our knowledge, this is the largest and most recent corpus so far. It was manually annotated for offensiveness, and fine-grained hate speech. In order to encourage comparisons between future studies, the providers of the dataset experimented with different transformer architectures and SVM to benchmark the dataset for detecting offense and hate speech to encourage comparisons between future studies. They fine-tuned mono- and multilingual transformer models using their training data. For the monolingual task, they utilised AraBERT and QARiB; for the multilingual models, they fine-tuned mBERT and XLM-RoBERTa. It was obvious in their reported results that monolingual models significantly outperformed the multilingual models. For offensive classification, the QARiB model achieved an F1-score equal to 82.31%, outperforming all other models, including AraBERT which came second with an F1-score equal to 80.02%. In contrast, AraBERT outperformed QARIB, achieving an F1-score of 80.14% and winning first place for hate speech detection.

III. DATASET PREPARATION

A. Dataset Description

We used the largest and most recently released dataset for offensive language and hate speech detection in Arabic [26]. The dataset consists of 12,698 tweets and defined according to the following: offensive language is a language containing any kind of impolite language such as insults, slurs, threats, and encouraging violence. Hate speech is any kind of offensive language that targets a person or group of people based on six common characteristics: disability, gender, ideology, race, religion, or social class. The task of offensive language detection was annotated using two labels OFF (offensive example) and NOT_OFF (not offensive example). Hate speech detection was annotated using two labels HS (hate speech example) and NOT_HS (not hate speech example). The fine-grained hate speech was annotated using 7 labels: HS1 (race/ethnicity/nationality), HS2 (f religion/belief), HS3 (ideology), HS4 (disability/disease), HS5 (social class), HS6 (gender), and NOT_HS (not hate-speech).

In our study, we passed the dataset through a set of preprocessing levels that started with cleaning the data to remove noise, followed by converting emoji into textual description, and then finding the sentiment of the tweet (i.e., positive, negative, and neutral) and appending the sentiment to the text of the tweet as additional textual features. Indeed, various levels of improved preprocessing have been examined when utilised with different techniques to identify their role in improving the performance of every built model.

B. Cleaning

The key component of any successful NLP application is to remove noise and reduce data sparsity as much as possible. It is well known that Arabic used in user-generated online content, including social media, is written in Arabic dialects which have many lexical, syntactic, and morphological differences, increasing the data sparsity of any corpus collected from online sources [52], [53]. In addition, online content is usually noisy, with a considerable number of tags, excessive spaces, repeated characters, Arabizi in which some people, when writing online, transliterate Arabic using Latin letters and numerals. In preprocessing step, we cleaned the text in order to reduce noise and data sparsity by the following:

- removing HTML tags and other symbols such as <LF>
- removing hashtags # and mentions @
- replacing underscore symbol _ of hastags into space
- removing URLs and retweets RT
- removing all types of diacritical marks, punctuation marks, mathematical signs and symbols
- removing repeated letters
- removing symbols different from emojis
- normalising different forms of alif into a bare alif (alif without hamzah), normalising taa' marbutah to haa' and normalising the dotless yaa' (alif maqsurah) to yaa'.

To perform normalisation and repeated letter removal, we used the AraNLP library [54].

C. Textual Emoji Description

An emoji is a pictograph embedded in text in electronic communication and web pages that conveys emotional cues, attitudes, and feelings that cannot be concluded from typed conversations. They exist in various forms such as facial expressions, common objects, animals, places, and types of weather. Thus, emojis can play an important role in the detection of offensive languages. In [55], the authors observed that the most frequent personal attack on Arabic Twitter is to call a person an animal names such as (*kalb*, “dog”)¹ and (*HmAr*, “donkey”). The same observation was reported in a previous study of [57]. In addition, some face emojis (anger and disgust) and objects (shoes) are widely used in offensive communication [57].

Converting an emoji to its textual description was one of the intensive preprocessing steps used to prepare the text before using it to train an SVM model in the study of [40]. Their SVM model achieved F1-score equal to 95% for detection of hate speech in Arabic text, ranking first in the shared task on Arabic offensive language detection in the OSACT4 workshop co-located with LREC 2020. We investigated the use of emoji descriptions as additional textual features that can be appended to tweets with the original text. Different settings can be examined, such as the technique utilised as well as the size and balance of the annotated examples for each class. We plan to investigate the importance of emojis themselves or their textual descriptions when used with deep contextualised word representation techniques, such as BERT, and compare it with other traditional ML techniques, such as SVM and logit.

We used “demoji” package in python [58] which accurately find emojis from a blob of text using data from the Unicode Consortium’s emoji code repository. After finding emojis in each tweet, we replaced them in the text with their code (i.e., textual description) equivalents. Fig. 1 shows the results of using the demoji package on a tweet from the training data. The results of the textual descriptions of emojis are in English. Thus, we utilised the Google Translate API to convert the textual descriptions from English to Arabic.

D. Sentiment Analysis

Sentiment analysis can generally be defined as the use of NLP techniques to detect, recognise, and quantify affective states and subjective information. However, the basic idea of

¹Throughout the paper, Arabic words are represented as follows: (HSB transliteration, ‘English gloss’). More details about the Habash–Soudi–Buckwalter (HSB) scheme can be found in [56]

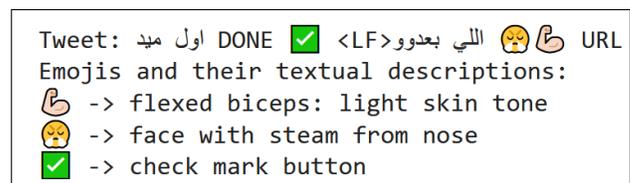


Fig. 1. Extracting Emojis form a Tweet and Finding their Descriptions.

A. BERT Model Classifier

BERT [65] is a multilayer bidirectional Transformer encoder based on the original implementation of transformer architecture introduced by Vaswani et al. [66]. The BERT model resulted in significant improvements in a considerable number of downstream tasks. Furthermore, a wide range of research works on Arabic hate speech and offensive language detection, including those participating in the 2020 shared task on Arabic offensive language detection, have proven its potential to handle the task [21]. In addition, the BERT-based model was the best benchmark system trained on the same dataset we employed in this study, allowing us to compare our proposed method with the best benchmark system.

In our study, we built a BERT-based model by fine-tuning AraBERT [67] on the training data. We selected “AraBERTv0.2-Twitter-base” variant of AraBERT that supports emojis and dialectal Arabic words. We also applied a segmentation function using Farasa to segment the text for the model. We built five BERT models for each task (i.e., offensive language detection, hate-speech detection, and fine-grained hate-speech detection), resulting in a total of 15 models. The five models were built using five versions of the dataset according to various levels of preprocessing, as previously illustrated in Table II.

B. SVM Classifier

We used word n -grams with n in the range [1, 3] weighted using Term Frequency-Inverse Document Frequency (TF-IDF). We also used character n -grams with n in the range [2, 5] only from the text inside word boundaries using token counts. The word-based TF-IDF vector and character-based count vector were used as features to train the SVM. As in the BERT model, we built 15 SVM classifiers to examine the various situations: three tasks in addition to the five levels of preprocessing the dataset to prepare it before building the models.

C. Logistic Regression Classifier

The same features used in the SVM were examined using a logistic regression classifier. Therefore, we used word n -grams with n in the range [1, 3] weighted using Term Frequency-Inverse Document Frequency (TF-IDF). We also used character n -grams with n in the range [2, 5] only from the text inside word boundaries using token counts. The word-based TF-IDF vector and character-based count vector were utilised as features to train the logistic regression classifier. The sklearn package in Python was used to train the classifier. The maximum number of iterations for logistic regression in the package was set to 100 by default. In our experiments, we increased this value to 800 iterations to obtain the trained model. Similar to the BERT model and SVM classifier, we built 15 logit classifiers.

V. EXPERIMENTAL SETUP

We used the same splits prepared by the data providers where the dataset was partitioned into three parts: 8,888 (70%) for training, 1,269 (10%) for development, and 2,541 (20%) for testing. Table III and Table IV show the distribution of offensive and hate speech data.

TABLE III. DISTRIBUTION OF OFFENSIVE AND HATE SPEECH DATA [26]

	Train	Dev	Test	Total
OFF	3,172	404	887	4,463
NOT_OFF	5,716	865	1,654	8,235
HS	959	109	271	1,339
NOT_HS	7,929	1,160	2,270	11,359
Total	8,888	1,269	2,541	12,698

TABLE IV. DISTRIBUTION OF FINE-GRAINED HATE SPEECH DATA. “N.A.” STANDS FOR NOT AVAILABLE

	Train	Dev	Test
HS1	260	28	N.A.
HS2	27	4	N.A.
HS3	144	14	N.A.
HS4	1	0	N.A.
HS5	72	10	N.A.
HS6	456	52	N.A.
NOT_HS	7928	1161	2,270
Total	8,888	1,269	2,541

At the time of writing the paper, the gold-standard labels for the training and development sets were publicly available, whereas only the tweets of the test set were available without the gold-standard labels. The providers of the data, however, accepted our request and helped us evaluate our best-performing model on the labelled test set and provided us with the results of our best performing model for all three tasks. Therefore, we utilised the results of our built models evaluated on development set in order to compare between them and to evaluate the various preprocessing settings we suggested in this paper. Next, the results of our best-performing model when evaluated on the test set were then compared with the benchmark systems [26] that were trained and tested using the same dataset we used in this study. The employed evaluation metrics in our study are macro-averaged precision, recall and F1- score, in addition to accuracy.

VI. RESULTS AND DISCUSSION

The results of the SVM classifiers evaluated on the development set for the three tasks are presented in Table V. These classifiers are trained on the versions of the dataset resulting from the five different levels of preprocessing: Orgi (original tweets), CLN (after cleaning tweets from noise), CLN+SA (after cleaning and appending the sentiment of the tweet to its text), CLN+EmoTxt (after cleaning and replacing emojis with their textual descriptions), and CLN+Emotxt+SA (after cleaning, replacing emojis with their textual descriptions, and appending the sentiment of the tweet to its text). We used the macro-averaged F1-score to rank the various classifiers for each task. For offensive language detection, we observe that the SVM classifier leads to the best results after cleaning the dataset, replacing emojis with their textual descriptions and appending the sentiment of each example to its text. For hate speech detection, the best performing SVM classifier is obtained using sentiment analysis, but without the need for emojis conversion. For fine-grained hate speech detection, it

TABLE V. ACCURACY AS WELL AS MACRO-AVERAGED (P)RECISION, (R)ECALL AND F1 SCORE OF SVM CLASSIFIERS ON DEVELOPMENT SET

Offensive language detection				
	Acc	P	R	F1
Orgi	79.84	78.81	72.48	74.25
CLN	80.00	78.78	72.92	74.64
CLN+SA	80.94	79.58	74.61	76.22
CLN+EmoTxt	80.39	79.45	73.28	75.07
CLN+EmoTxt+SA	81.26	80.54	74.44	76.28
Hate speech detection				
	Acc	P	R	F1
Orgi	92.44	85.48	58.04	61.64
CLN	92.68	85.78	59.83	64.12
CLN+SA	92.83	86.57	60.75	65.37
CLN+EmoTxt	92.68	85.78	59.83	64.12
CLN+EmoTxt+SA	92.83	87.61	60.34	64.89
Fine-grained hate speech detection				
	Acc	P	R	F1
Orgi	92.13	37.50	18.57	20.23
CLN	92.36	37.92	19.63	21.79
CLN+SA	92.28	37.72	19.35	21.48
CLN+EmoTxt	92.13	23.41	18.33	19.56
CLN+EmoTxt+SA	92.13	23.41	18.33	19.56

TABLE VI. ACCURACY AS WELL AS MACRO-AVERAGED (P)RECISION, (R)ECALL AND F1 SCORE OF LOGISTIC REGRESSION CLASSIFIERS ON DEVELOPMENT SET

Offensive language detection				
	Acc	P	R	F1
Orgi	78.66	76.77	71.48	73.01
CLN	79.61	77.90	72.83	74.39
CLN+SA	80.55	78.53	74.84	76.16
CLN+EmoTxt	79.61	77.66	73.16	74.62
CLN+EmoTxt+SA	81.10	79.69	74.92	76.50
Hate speech detection				
	Acc	P	R	F1
Orgi	93.07	87.59	62.13	67.18
CLN	93.07	86.72	62.54	67.61
CLN+SA	93.15	87.90	62.19	67.77
CLN+EmoTxt	93.07	88.59	61.71	66.74
CLN+EmoTxt+SA	92.97	88.72	61.78	66.90
Fine-grained hate speech detection				
	Acc	P	R	F1
Orgi	91.99	51.97	22.34	23.91
CLN	92.76	53.42	22.03	25.25
CLN+SA	92.76	52.71	21.77	24.39
CLN+EmoTxt	92.77	52.86	21.75	24.87
CLN+EmoTxt+SA	92.91	52.86	21.75	24.87

is sufficient to clean the dataset from noises before training the SVM classifier in order to obtain the best results of the algorithm for the task.

The results of the logistic regression (logit) classifiers for the three tasks are presented in Table VI. The logit classifier for hate speech detection achieves the best performance when using sentiment analysis as well as cleaning noisy data, yielding an F1-score of 67.77. Both SVM and logit classifiers require the same preprocessing level (CLN+SA) to achieve the best performance. For offensive language detection, the use of emojis conversion and sentiment analysis when preparing the dataset plays an important role in obtaining the best performing logit classifier, achieving an F1-score equal to 76.50 (slightly better than the best SVM classifier for the same task which achieves F1-score = 76.28). For fine-grained hate speech detection, the only required preprocessing of the data to build a logit classifier with the best performance is to clean the text from noise. In fact, the use of sentiments as additional features or converting emojis into their textual code leads to a decline in the performance of the built classifier. This observation matches what we noticed with the SVM classifiers for fine-grained hate speech detection.

The BERT models outperformed all other SVM and logistic regression classifiers regardless of the preprocessing level used to prepare the dataset, as shown in Table VII. The best performing BERT models achieved an F1-score equal to 85.93, 81.89, and 48.72 for offensive language detection, hate speech detection, and fine-grained hate speech detection, respectively. Regarding the optimal preprocessing steps that can be applied to the dataset before training to improve the model's performance regardless of the utilised ML technique, we observe

that CLN+EmoTxt+SA (i.e., cleaning the data, converting emojis to textual words, and appending sentiments as additional textual features) always improves the performance of the model for the offensive language detection task. This can be attributed to the fact that offensive language detection is considered easier than detecting hate speech or identifying the exact type of hate speech. In offensive language detection, every tweet that contains an impolite language, including hate speech, is considered offensive language according to the annotation guidelines followed by the providers of the dataset [26]. Therefore, adding additional features, such as textual descriptions of emojis or sentiments, to each tweet will confirm the boundaries that should be learned by the algorithms to distinguish between normal and offensive tweets. That is, almost all offensive tweets have the sentiment "negative" added as additional features, while normal tweets usually have "positive" or "neutral" sentiments added as additional features. Emojis on the other hand are also considered distinctive features in the case of offensive language detection as offensive emojis that express anger and disgust are not commonly found in normal tweets (i.e., not offensive). Thus, converting emojis to textual descriptions increases the number of additional distinctive features. That is, instead of having only one emoji, we will have, by converting emoji to description, a phrase with words expressing anger and disgust. In contrast, the CLN+EmoTxt+SA usually has a fluctuating impact on the performance of the model for the task of hate speech detection, as it sometimes slightly improves the performance of the model as seen in Table V and Table VII, and sometimes deteriorates the performance of the model, as shown in Table VI. However, CLN+EmoTxt+SA did not allow the model to achieve the

TABLE VII. ACCURACY AS WELL AS MACRO-AVERAGED (P)RECISION, (R)ECALL AND F1 SCORE OF BERT MODELS ON **DEVELOPMENT SET**

Offensive language detection				
	Acc	P	R	F1
Orgi	86.77	48.51	85.55	84.98
CLN	87.63	85.77	85.72	85.74
CLN+SA	87.72	85.93	85.63	85.79
CLN+EmoTxt	87.95	86.51	85.36	85.87
CLN+EmoTxt+SA	87.80	86.52	85.40	85.93
Hate speech detection				
	Acc	P	R	F1
Orgi	93.88	83.60	79.41	80.91
CLN	94.57	82.62	81.90	81.76
CLN+SA	94.33	82.67	80.34	81.89
CLN+EmoTxt	94.09	81.03	81.81	81.41
CLN+EmoTxt+SA	94.09	80.85	82.63	81.71
Fine-grained hate speech detection				
	Acc	P	R	F1
Orgi	92.99	47.68	46.78	45.79
CLN	93.54	49.74	49.13	48.72
CLN+SA	93.62	49.46	49.15	48.55
CLN+EmoTxt	93.31	48.91	48.59	47.91
CLN+EmoTxt+SA	93.07	48.02	47.82	41.16

best performance. This can be attributed to the fact that hate speech in the utilised dataset is considered a type of offensive language. That is, a tweet may contain impolite language and is considered offensive but not “hate-speech”. Also, the general unbalance between various classes in the dataset, as seen in Table III, is more obvious in the case of hate speech as there are few annotated hate speech examples compared to not-hate-speech examples. Therefore, converting emojis to textual descriptions actually increases data sparsity and cannot be seen as a vital preprocessing step in the case of hate speech detection, as seen in Tables V, VI, and VII. In the case of fine-grained hate speech detection, there are six types of hate speech, and there is a huge unbalance between the number of annotated examples for these types. For example, we have 260 tweets labelled as “HS1” (race/ethnicity/nationality), 27 tweets were annotated as “HS2” (religion/belief), and 7,928 were annotated as not-hate-speech. Therefore, the preprocessing step of emoji conversion does not have a significant positive effect on the overall performance of the model, as it may increase the data sparsity, especially for unbalanced datasets with few annotated examples and overlapping classes. However, we observe that cleaning the data improves the performance of the model, regardless of the utilised training algorithm. In addition, the use of sentiments as additional features appended to the tweet’s text has a good impact on the overall performance of the model. This positive impact is less obvious when the number of annotated examples is insufficient, to distinguish between a large number of overlapping classes, such as in the case of fine-grained hate speech detection.

The best performing model (BERT model) for each task was selected to be evaluated on the test part of the dataset. The results are presented in Table VIII, which shows that

TABLE VIII. PERFORMANCE COMPARISON OF OUR BEST PERFORMING MODEL AND FOUR BENCHMARK SYSTEMS [26] ON **TEST SET**

Offensive language detection				
	Acc	P	R	F1
AraBERT	92.64	81.04	79.31	80.14
QARiB	92.99	82.99	77.72	80.04
mBERT	91.26	77.55	73.34	75.20
XLM-RoBERTa	92.29	79.96	78.79	79.36
Our Model_{best}	93.30	83.00	80.70	81.80
Hate speech detection				
	Acc	P	R	F1
AraBERT	82.09	80.50	79.63	80.02
QARiB	84.02	82.53	82.11	82.31
mBERT	76.43	74.09	73.32	73.66
XLM-RoBERTa	75.00	72.50	72.47	72.48
Our Model_{best}	85.90	84.60	84.10	84.30

our best performing model with its suggested preprocessing levels outperforms all other benchmark models for two tasks: offensive language detection and hate speech detection [26]. The study in [26] did not provide a benchmark system for fine-grained hate speech detection. However, our proposed BERT-based model achieved an F1-score equal to 45.10% on the test set. The precision, recall, and accuracy were 48.20%, 46.10%, and 92.10% respectively.

VII. CONCLUSION

In this study, we proposed an automatic method for detecting offensive language and fine-grained hate speech from Arabic tweets. We compared BERT with two conventional machine learning techniques (SVM and logistic regression). We also investigated the use of sentiments and textual descriptions of emojis as appending features in the dataset, along with the textual content of the tweets. The experiments clarified that the BERT-based model results in the best performance, surpassing the best benchmark systems in the literature, for all three tasks: (a) offensive language detection with an 84.3% F1-score, (b) hate speech detection with an 81.8% F1-score, and (c) fine-grained hate-speech recognition (e.g., race, religion, social class, etc.) with a 45.1% F1-score. Analysing the sentiment of each tweet and using it as a feature slightly improves the performance of the models when detecting offensive language and hate speech, but has little positive effect on the performance of models for recognising the type of hate speech. The use of textual emoji descriptions as features can improve or deteriorate the performance of the models depending on the size of the annotated examples per class and whether the emojis are considered distinctive features between classes. That is, when the number of annotated examples is limited while the classes overlap in the feature space, emojis may not be considered as distinctive features, and converting them to their textual description as additional features may increase the data sparsity and, therefore, deteriorate the performance of the model. However, our proposed models and various levels of preprocessing lead to better results than the benchmark systems reported in the previous study.

REFERENCES

- [1] G. S. O’Keeffe, K. Clarke-Pearson *et al.*, “The impact of social media on children, adolescents, and families,” *Pediatrics*, vol. 127, no. 4, pp. 800–804, 2011.
- [2] E. R. Munro, “The protection of children online: a brief scoping review to identify vulnerable groups,” *Childhood Wellbeing Research Centre*, 2011.
- [3] P. Burnap, O. Rana, M. Williams, W. Housley, A. Edwards, J. Morgan, L. Sloan, and J. Conejero, “Cosmos: Towards an integrated and scalable service for analysing social media on demand,” *International Journal of Parallel, Emergent and Distributed Systems*, vol. 30, no. 2, pp. 80–100, 2015.
- [4] Z. Waseem, T. Davidson, D. Warmesley, and I. Weber, “Understanding abuse: A typology of abusive language detection subtasks,” in *Proceedings of the First Workshop on Abusive Language Online*, 2017, pp. 78–84.
- [5] 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, April 3, 2017, Valencia, Spain*. Association for Computational Linguistics, 2019, pp. 1–10.
- [6] P. Mishra, H. Yannakoudakis, and E. Shutova, “Tackling online abuse: A survey of automated abuse detection methods,” *arXiv preprint arXiv:1908.06024*, 2019.
- [7] M. Wiegand, M. Geulig, and J. Ruppenhofer, “Implicitly abusive comparisons—a new dataset and linguistic analysis,” in *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, 2021, pp. 358–368.
- [8] B. Vidgen, T. Thrush, Z. Waseem, and D. Kiela, “Learning from the worst: Dynamically generated datasets to improve online hate detection,” in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 2021, pp. 1667–1682.
- [9] R. Hada, S. Sudhir, P. Mishra, H. Yannakoudakis, S. Mohammad, and E. Shutova, “Ruddit: Norms of offensiveness for english reddit comments,” in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 2021, pp. 2700–2717.
- [10] P. Röttger, B. Vidgen, D. Nguyen, Z. Waseem, H. Margetts, and J. Pierrehumbert, “Hatecheck: Functional tests for hate speech detection models,” in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 2021, pp. 41–58.
- [11] R. Kumar, A. K. Ojha, M. Zampieri, and S. Malmasi, Eds., *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*. Santa Fe, New Mexico, USA: Association for Computational Linguistics, 2018. [Online]. Available: <https://aclanthology.org/W18-4400>
- [12] R. Kumar, A. K. Ojha, S. Malmasi, and M. Zampieri, “Benchmarking aggression identification in social media,” in *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*. Santa Fe, New Mexico, USA: Association for Computational Linguistics, Aug. 2018, pp. 1–11. [Online]. Available: <https://aclanthology.org/W18-4401>
- [13] Z. Waseem, W. H. K. Chung, D. Hovy, and J. Tetreault, Eds., *Proceedings of the First Workshop on Abusive Language Online*. Vancouver, BC, Canada: Association for Computational Linguistics, Aug. 2017. [Online]. Available: <https://aclanthology.org/W17-3000>
- [14] D. Fišer, R. Huang, V. Prabhakaran, R. Voigt, Z. Waseem, and J. Wernimont, Eds., *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*. Brussels, Belgium: Association for Computational Linguistics, Oct. 2018. [Online]. Available: <https://aclanthology.org/W18-5100>
- [15] S. T. Roberts, J. Tetreault, V. Prabhakaran, and Z. Waseem, Eds., *Proceedings of the Third Workshop on Abusive Language Online*. Florence, Italy: Association for Computational Linguistics, Aug. 2019. [Online]. Available: <https://aclanthology.org/W19-3500>
- [16] S. Akiwowo, B. Vidgen, V. Prabhakaran, and Z. Waseem, Eds., *Proceedings of the Fourth Workshop on Online Abuse and Harms*. Online: Association for Computational Linguistics, Nov. 2020. [Online]. Available: <https://aclanthology.org/2020.alw-1.0>
- [17] A. Mostafazadeh Davani, D. Kiela, M. Lambert, B. Vidgen, V. Prabhakaran, and Z. Waseem, Eds., *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*. Online: Association for Computational Linguistics, Aug. 2021. [Online]. Available: <https://aclanthology.org/2021.woah-1.0>
- [18] Z. Waseem, B. Vidgen, L. Mathias, and A. Davani, “The 6th Workshop on Online Abuse and Harms (2022),” 2022, [Online]. Available: <https://www.workshoponlineabuse.com/> (accessed 20-April-2022).
- [19] M. Zampieri, S. Malmasi, P. Nakov, S. Rosenthal, N. Farra, and R. Kumar, “Semeval-2019 task 6: Identifying and categorizing offensive language in social media (offenseval),” in *Proceedings of the 13th International Workshop on Semantic Evaluation*, 2019, pp. 75–86.
- [20] M. Zampieri, P. Nakov, S. Rosenthal, P. Atanasova, G. Karadzhov, H. Mubarak, L. Derczynski, Z. Pitenis, and Ç. Çöltekin, “SemEval-2020 task 12: Multilingual offensive language identification in social media (OffensEval 2020),” in *Proceedings of the Fourteenth Workshop on Semantic Evaluation*. Barcelona (online): International Committee for Computational Linguistics, Dec. 2020, pp. 1425–1447. [Online]. Available: <https://aclanthology.org/2020.semeval-1.188>
- [21] H. Mubarak, K. Darwish, W. Magdy, T. Elsayed, and H. Al-Khalifa, “Overview of OSACT4 Arabic offensive language detection shared task,” in *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*. Marseille, France: European Language Resource Association, May 2020, pp. 48–52. [Online]. Available: <https://aclanthology.org/2020.osact-1.7>
- [22] H. Mubarak, “Arabic Hate Speech 2022 Shared Task!” 2022, [Online]. Available: <https://sites.google.com/view/arabichate2022/home> (accessed 20-April-2022).
- [23] T. Davidson, D. Warmesley, M. Macy, and I. Weber, “Automated hate speech detection and the problem of offensive language,” in *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 11, no. 1, 2017, pp. 512–515.
- [24] P. Fortuna and S. Nunes, “A survey on automatic detection of hate speech in text,” *ACM Computing Surveys (CSUR)*, vol. 51, no. 4, pp. 1–30, 2018.
- [25] Z. Waseem and D. Hovy, “Hateful symbols or hateful people? predictive features for hate speech detection on twitter,” in *Proceedings of the NAACL student research workshop*, 2016, pp. 88–93.
- [26] H. Mubarak, S. Hassan, and S. A. Chowdhury, “Emojis as anchors to detect arabic offensive language and hate speech,” *arXiv preprint arXiv:2201.06723*, 2022.
- [27] P. Burnap and M. L. Williams, “Cyber hate speech on twitter: An application of machine classification and statistical modeling for policy and decision making,” *Policy & internet*, vol. 7, no. 2, pp. 223–242, 2015.
- [28] M. Dadvar, D. Trieschnigg, and F. d. Jong, “Experts and machines against bullies: A hybrid approach to detect cyberbullies,” in *Canadian conference on artificial intelligence*. Springer, 2014, pp. 275–281.
- [29] M. Dadvar, D. Trieschnigg, R. Ordelman, and F. d. Jong, “Improving cyberbullying detection with user context,” in *European Conference on Information Retrieval*. Springer, 2013, pp. 693–696.
- [30] B. Gambäck and U. K. Sikdar, “Using convolutional neural networks to classify hate-speech,” in *Proceedings of the first workshop on abusive language online*, 2017, pp. 85–90.
- [31] A. Kontostathis, K. Reynolds, A. Garron, and L. Edwards, “Detecting cyberbullying: query terms and techniques,” in *Proceedings of the 5th annual acm web science conference*, 2013, pp. 195–204.
- [32] J. Pavlopoulos, P. Malakasiotis, and I. Androutsopoulos, “Deep learning for user comment moderation,” *arXiv preprint arXiv:1705.09993*, 2017.
- [33] H.-P. Su, Z.-J. Huang, H.-T. Chang, and C.-J. Lin, “Rephrasing profanity in chinese text,” in *Proceedings of the First Workshop on Abusive Language Online*, 2017, pp. 18–24.

- [34] D. Fišer, T. Erjavec, and N. Ljubešić, "Legal framework, dataset and annotation schema for socially unacceptable online discourse practices in slovene," in *Proceedings of the first workshop on abusive language online*, 2017, pp. 46–51.
- [35] N. Ljubešić, T. Erjavec, and D. Fišer, "Datasets of slovene and croatian moderated news comments," in *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, 2018, pp. 124–131.
- [36] I. Clarke and J. Grieve, "Dimensions of abusive language on twitter," in *Proceedings of the first workshop on abusive language online*, 2017, pp. 1–10.
- [37] E. A. Abozinadah, A. V. Mbaziira, and J. Jones, "Detection of abusive accounts with arabic tweets," *International Journal of Knowledge Engineering-IACSIT*, vol. 1, no. 2, pp. 113–119, 2015.
- [38] H. Mubarak, K. Darwish, and W. Magdy, "Abusive language detection on arabic social media," in *Proceedings of the first workshop on abusive language online*, 2017, pp. 52–56.
- [39] A. Alakrot, L. Murray, and N. S. Nikolov, "Dataset construction for the detection of anti-social behaviour in online communication in arabic," *Procedia Computer Science*, vol. 142, pp. 174–181, 2018.
- [40] F. Husain, "Osact4 shared task on offensive language detection: Intensive preprocessing-based approach," in *LREC 2020 Workshop Language Resources and Evaluation Conference 11–16 May 2020*, p. 53.
- [41] A. I. Alharbi and M. Lee, "Combining character and word embeddings for the detection of offensive language in arabic," in *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, 2020, pp. 91–96.
- [42] A. Alakrot, L. Murray, and N. S. Nikolov, "Towards accurate detection of offensive language in online communication in arabic," *Procedia computer science*, vol. 142, pp. 315–320, 2018.
- [43] H. Haddad, H. Mulki, and A. Oueslati, "T-hsab: A tunisian hate speech and abusive dataset," in *International Conference on Arabic Language Processing*. Springer, 2019, pp. 251–263.
- [44] H. Mulki, H. Haddad, C. B. Ali, and H. Alshabani, "L-hsab: A levantine twitter dataset for hate speech and abusive language," in *Proceedings of the third workshop on abusive language online*, 2019, pp. 111–118.
- [45] A. Abuzayed and T. Elsayed, "Quick and simple approach for detecting hate speech in arabic tweets," in *Proceedings of the 4th workshop on open-source Arabic Corpora and processing tools, with a shared task on offensive language detection*, 2020, pp. 109–114.
- [46] A. Keleg, S. R. El-Beltagy, and M. Khalil, "Asu_opto at osact4-offensive language detection for arabic text," in *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, 2020, pp. 66–70.
- [47] H. Mubarak and K. Darwish, "Arabic offensive language classification on twitter," in *International Conference on Social Informatics*. Springer, 2019, pp. 269–276.
- [48] H. Mohaouchane, A. Mourhir, and N. S. Nikolov, "Detecting offensive language on arabic social media using deep learning," in *2019 sixth international conference on social networks analysis, management and security (SNAMS)*. IEEE, 2019, pp. 466–471.
- [49] N. Albadi, M. Kurdi, and S. Mishra, "Investigating the effect of combining gru neural networks with handcrafted features for religious hatred detection on arabic twitter space," *Social Network Analysis and Mining*, vol. 9, no. 1, pp. 1–19, 2019.
- [50] A. Elmadany, C. Zhang, M. Abdul-Mageed, and A. Hashemi, "Leveraging affective bidirectional transformers for offensive language detection," in *LREC 2020 Workshop Language Resources and Evaluation Conference 11–16 May 2020*, p. 102.
- [51] M. Djandji, F. Baly, W. Antoun, and H. Hajj, "Multi-task learning using arabert for offensive language detection," in *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, 2020, pp. 97–101.
- [52] M. J. Althobaiti, "Creation of annotated country-level dialectal arabic resources: An unsupervised approach," *Natural Language Engineering*, pp. 1–42, 2021, DOI: 10.1017/s135132492100019x.
- [53] —, "Automatic arabic dialect identification systems for written texts: a survey," *arXiv preprint arXiv:2009.12622*, 2020.
- [54] M. Althobaiti, U. Kruschwitz, and M. Poesio, "ArANLP: a Java-based Library for the Processing of Arabic Text," in *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Association for Computational Linguistics. Reykjavik, Iceland: European Language Resources Association (ELRA), 2014, pp. 4134–4138.
- [55] H. Mubarak, A. Rashed, K. Darwish, Y. Samih, and A. Abdelali, "Arabic offensive language on twitter: Analysis and experiments," in *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, 2021, pp. 126–135.
- [56] N. Habash, A. Soudi, and T. Buckwalter, "On arabic transliteration," in *Arabic computational morphology*. Springer, 2007, pp. 15–22.
- [57] S. A. Chowdhury, H. Mubarak, A. Abdelali, S.-g. Jung, B. J. Jansen, and J. Salminen, "A multi-platform arabic news comment dataset for offensive language detection," in *Proceedings of the 12th Language Resources and Evaluation Conference*, 2020, pp. 6203–6212.
- [58] B. Solomon, "demoji 1.1.0," 2021, [Online]. Available: <https://pypi.org/project/demoji/> (accessed 20-April-2022).
- [59] L. Zhang, S. Wang, and B. Liu, "Deep learning for sentiment analysis: A survey," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 8, no. 4, p. e1253, 2018.
- [60] M. Al-Ayyoub, A. A. Khamaiseh, Y. Jararweh, and M. N. Al-Kabi, "A comprehensive survey of arabic sentiment analysis," *Information processing & management*, vol. 56, no. 2, pp. 320–342, 2019.
- [61] G. Inoue, B. Alhafni, N. Baimukan, H. Bouamor, and N. Habash, "The interplay of variant, size, and task type in Arabic pre-trained language models," in *Proceedings of the Sixth Arabic Natural Language Processing Workshop*. Kyiv, Ukraine (Online): Association for Computational Linguistics, Apr. 2021.
- [62] M. Nabil, M. Aly, and A. Atiya, "Astd: Arabic sentiment tweets dataset," in *Proceedings of the 2015 conference on empirical methods in natural language processing*, 2015, pp. 2515–2519.
- [63] A. Elmadany, H. Mubarak, and W. Magdy, "Arsas: An arabic speech-act and sentiment corpus of tweets," *OSACT*, vol. 3, p. 20, 2018.
- [64] S. Rosenthal, N. Farra, and P. Nakov, "Semeval-2017 task 4: Sentiment analysis in twitter," in *Proceedings of the 11th international workshop on semantic evaluation (SemEval-2017)*, 2017, pp. 502–518.
- [65] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of NAACL-HLT*, 2019, pp. 4171–4186.
- [66] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, E. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.
- [67] W. Antoun, F. Baly, and H. Hajj, "Arabert: Transformer-based model for arabic language understanding," in *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, 2020, pp. 9–15.

A Survey of Sink Mobility Models to Avoid the Energy-Hole Problem in Wireless Sensor Networks

Ghada Al-Mamari, Fatma Bouabdallah, Asma Cherif
Faculty of Computing and
Information Technology
Information Technology Dept.
King Abdulaziz University
Jeddah, Saudi Arabia

Abstract—Wireless Sensor Networks (WSN) refer to networks where the sensors are deployed in an environment to sense and select data. WSN sensor nodes have limited power and cannot be recharged easily. Consequently, the faster sensor nodes to deplete their energy budget are those close to the sink as they have to relay all data emanating from any sensor in the network. Thus, a hole of energy around the sink is created as the sink coverage nodes have drained their initial energy thus leading to sink unreachability. The WSN lifetime maximization problem has always been a hot research topic. Collecting data in WSN using a mobile sink is an efficient approach for achieving WSN longevity and preventing the energy hole problem. However, finding the optimal trajectory along with its appropriate flow routing is a challenging problem since many constraints should be considered. This paper discusses and compares several existing WSN-lifetime-maximization using sink mobility solutions. These solutions are mainly classified into two types: Linear Programming and Artificial Intelligence-based solutions. The state-of-the-art solutions are compared in terms of network topology, sojourn points and duration, buffer size, and overhearing. Finally, a discussion of the WSN lifetime maximization constraints is provided to define a promising sink mobility model.

Keywords—Energy hole problem; mobile sink; wireless sensor networks; linear programming; artificial intelligence

I. INTRODUCTION

Now-a-days, Wireless Sensor Networks (WSN) are used to sense and collect data from many environments that are not easy to reach. WSN consists of: (1) a sink node, which oversees collected data from the wireless sensor nodes for further processing, and (2) many sensor nodes usually scattered in a harsh environment to collect data and deliver it to the sink to serve a given application such as environmental monitoring [1] and military applications [2]. Wireless Sensor nodes rely on batteries power. Once they are scattered, it is hard to replace or recharge the batteries. Therefore, maximizing the WSN lifetime by optimizing nodes' energy is crucial.

The sensor nodes close to the sink are always responsible for forwarding other distant nodes' data to the sink. Consequently, due to this burdening, they are usually the earliest nodes that deplete their energy, while the other sensor nodes still have a good amount of energy [3]. Furthermore, if the sensors that are closer to the sink drain their energy budget, they create an energy hole around the sink. Fig. 1 illustrates the energy hole problem caused by the batteries' depletion of sensor nodes within the sink coverage (nodes with a red

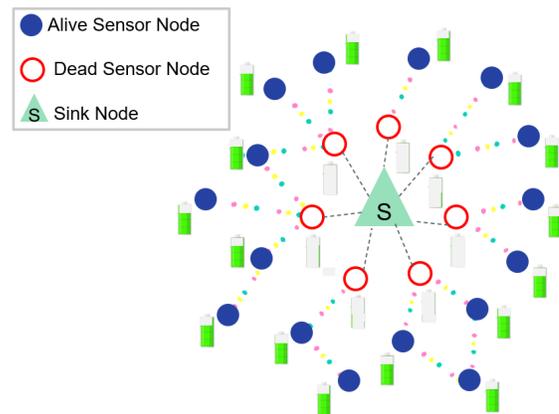


Fig. 1. Wireless Sensor Network and the Energy Hole Problem

border). If this energy hole occurs, the sink will become unreachable, and hence, the network is no longer operable as it won't be able to deliver the sensed data to the sink for further processing. To overcome this issue, many approaches have been proposed such as assigning sensor nodes multiple transmission power, nonuniform initial energy budget distribution, and sink mobility [4] (see in Section II). This survey focuses on maximizing WSN by exploiting sink mobility. Thus, the sink can move and stop at many points in the network to receive sensed data. Besides the sink mobility, the appropriate flow routing is defined to balance the energy consumption between sensor nodes.

This work considers only the algorithms that apply to delay-tolerant networks. This means that, when the sink is at location i , only the sensor nodes within the transmission range of the sink at this location can transmit their data, while the others have to delay their transmissions. So, in one round the sink has to collect data from all the nodes and returns back to its first location in D time which is the maximum delay-tolerant time [5].

The remainder of the survey is organized as follow: Section II presents some proposed approaches to overcome the energy hole problem, and the main constraints that are considered in WSN. Section III, IV and V discuss the most cited linear programming-based (LP-based), artificial intelligence-based (AI-based) and other solutions that are neither LP-based nor AI-based for maximizing the WSN lifetime using

a single mobile sink. Moreover, a comparative study among these existing WSN-lifetime maximization approaches will be conducted in Section VI. Finally, in section VII, we discuss some additional constraints that impact the WSN lifetime in order to define an optimal mobile sink solution that maximizes the network lifetime and other future research directions.

II. MITIGATING ENERGY HOLE PROBLEM APPROACHES

WSN nodes have limited resources (e.g., transmission range, buffer size, and batteries). Due to the battery constraint, the sink coverage nodes are the first nodes that deplete their energy due to the burdening of the relaying task from the distant nodes, causing the energy hole problem. Therefore, the energy hole problem is addressed in many research works. Many approaches had been proposed to mitigate the energy hole problem as shown in Fig. 2 such as: assigning sensor nodes multiple transmission power, nonuniform initial energy budget distribution, and sink mobility [4].

First, in the multiple transmission power approaches, to balance the energy consumption among sensor nodes in the WSN, two or three hops away sensors can send data to the sink directly by passing one or two hops away sensors. As a result, the sink coverage nodes are highly alleviated from the relaying task, and hence the WSN lifetime is improved. [6] and [7] proposed two different routing protocols that both aim at allowing the sensor nodes to either send a message to its immediate neighbors or send it directly to the sink node. Simulation results show that both protocols achieve longer network lifetime especially compared to the nominal transmission range solution. Regarding the nonuniform initial energy budget distribution approach, the sensor nodes that are closer to the sink are given a higher amount of energy than distant nodes. So, the sink coverage nodes will not die quickly because they have a higher energy budget. Nevertheless, their lifetime is always constrained as they have to forward all the remaining nodes data messages. Finally, the third approach considers a mobile sink that moves between the sensor nodes to relieve the sink coverage nodes from their heavy relaying task and aims at balancing the energy consumption among WSN sensor nodes such that every node will drain its energy budget smoothly and uniformly throughout the network. In our survey, we focus on work related to exploiting the sink mobility for

distribute the energy consumption hence maximizing the WSN lifetime.

It is worth noting that most of the reviewed papers agreed on three main constraints that describe the behavior of sensors and the sink in WSN, these constraints must be fulfilled to make the WSN work correctly. The constraints for the sensor node were defined as follows:

- The energy constraint: the total energy consumption at a node i due to the reception and transmission of data over the network lifetime T must not exceed its initial energy.
- The flow conservation law between sensor nodes: for any node, the sum of total incoming flow rates plus the self-generated data rate must be equal to the sum of the resulting outgoing flow rates at time t

As for the constraint related to the sink, it is defined as follows:

- The total traffic going into a sink in a duration of time must be equal to the amount of total generated data from all sensor nodes in that duration.

III. LINEAR PROGRAMMING-BASED SINK MOBILITY MODELS

After extensive research, we end up reviewing many papers that have proposed different sink mobility models using different approaches. We classified these research works as Linear Programming-based (LP-based), Artificial Intelligence-based (AI-based), and other sink mobility models as shown in Fig. 3.

In this section we will start discussing the LP-based models.

A. Exploiting a Single Mobile Sink

Many researchers proposed WSN maximization models that used a single mobile sink to collect data from WSN sensor nodes. These models are classified based on the decision-making placement as centralized and distributed solutions. They also vary based on their movement as discrete and semi-continuous.

Y. Yun et al. [8] formulated a distributed algorithm for the WSN-lifetime maximization problem. First, they convert the WSN lifetime maximization problem into a network flow problem on an expanded graph. For each possible sink location, correspond an expanded graph G_l , and a set of coverage nodes are defined. A path links each node in G_l to the sink. For each G_l , the sink takes T time to collect data from all sensors in that graph. So, the problem of WSN lifetime maximization is to maximize the number of sink tours T . The authors formulated the main three constraints of WSN as a linear problem to minimize the WSN maximum energy consumption. Then, they divided the problem into sub-problems using the fractional knapsack problem. Thus, they develop a distributed algorithm to decide at each node whether data should be transmitted to the sink or buffered locally. The distribution of the algorithm allows the solution to be built-in network protocols, and the decisions can be made at each sensor. The authors proved

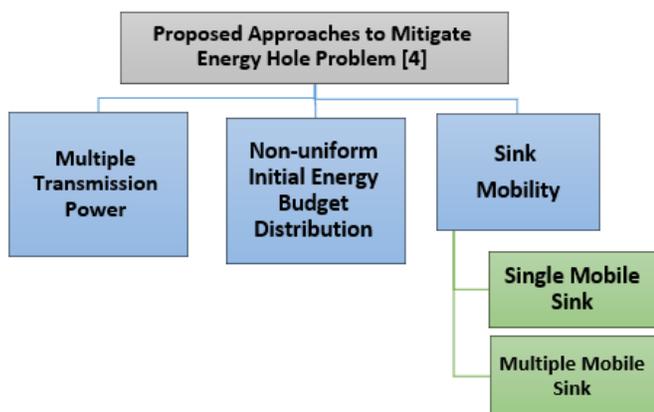


Fig. 2. Proposed Approaches to Mitigate the Energy Hole Problem

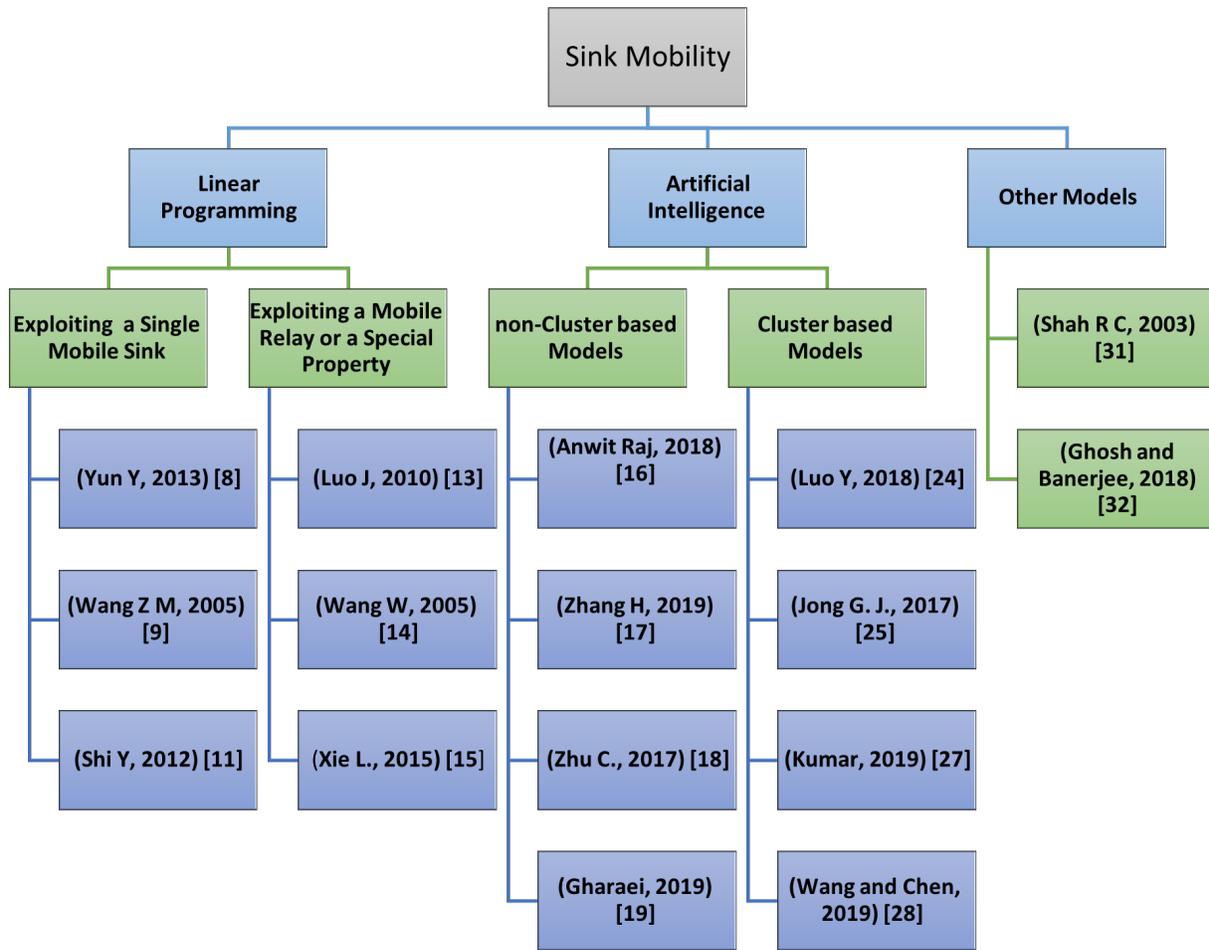


Fig. 3. Taxonomy of Reviewed Sink Mobility Models.

the efficiency and convergence to the optimal value. Moreover, they measured the Lyapunov drift at every iteration and proved that the queue size is bounded from above, which is equal to the maximum delay time D , and the average value of the total queue size has increased along with iterations. As a result, the increasing D will increase $T \times D$, which is the WSN lifetime.

On the other hand, many centralized sink mobility models were proposed. For instance, [9] proposed a joint algorithm to determine the sink movement and sojourn duration in a grid network topology. The suggested algorithm states that, when a sensor node is located at the same vertical or horizontal line as the sink, there is a single unique path between them which is the direct one. Otherwise, only the two routes are considered with equal hop count revealed by the rectangle boundary defined by a sensor i and the sink. To generate the energy model of their algorithm, a pair of horizontal and vertical dotted lines to enclose the nodes associated with the column and the row of the sink in the network is defined resulting in eight sub-areas of the grid as shown in Fig. 4: UL (Upper Left), UR (Upper Right), HL (Horizontal Left), HR (Horizontal Right), VA (Vertical Above), VB (Vertical Below), LL (Lower Left), LR (Lower Right).

Given a grid topology of size L (L being equal to the square root of the sensor nodes number), a sensor node i , its

sub-area, and position (i.e., column and row indexes (x, y)), the energy consumption is calculated as follows:

$$C_i^k \begin{cases} er[(x+1)(1+L)-1], & \text{if } i \in HL \\ er[(L-x)(1+L)-1], & \text{if } i \in HR \\ er[(y+1)(1+L)-1], & \text{if } i \in VA \\ er[(L-y)(1+L)-1], & \text{if } i \in VB \\ er(1+x+y), & \text{if } i \in UL \\ er(L-x+y), & \text{if } i \in UR \\ er(L+x-y), & \text{if } i \in LL \\ er(2L-x-y-1), & \text{if } i \in LR \\ er[(x+1)(1+L)-1], & \text{if } i \in HL \\ er, & \text{if } i = k \end{cases}$$

To find the optimal sojourn times at the grid cells that maximize the network lifetime, this energy model is converted to a linear programming model as follows:

$$\begin{aligned} \text{Max } z &= \sum_{k \in N} t_k \\ \text{Such that } \sum_{k \in N} C_i^k \times t_k &\leq e_0, & K \in i \\ t_k &\geq 0, & K \in N \end{aligned}$$

where the WSN lifetime denoted by z is the sum of all sink sojourn duration t_k (in seconds) at all sojourn points k , e_0 is the node's initial energy (in Joules), and C_i^k is the energy consumption of data transmitted in node i while the sink sojourns at position k (in Joules/bit).

Finally, the authors implemented their solution using LINGO [10] and provided visualization of the sink sojourn duration and the nodes' energy consumption during the network lifetime. The results showed the sink sojourn duration follow a pattern that is distributed between the four corner nodes of the two-dimensional grid. This pattern achieves a balanced energy consumption among sensor nodes, thus maximizing the WSN lifetime. As a limitation, there is a computational overhead for finding the optimal sojourn points and duration.

It is worth noting that the previously-mentioned solutions propose discrete mobility models. Other works such as [11] and [12] proposed a semi-continuous novel flow routing solution to solve the Unconstrained-Mobile Base station (U-MB) problem, wherein the base station is continuously roaming anywhere in the WSN environment. This proposed algorithm is driven from the Constrained-Mobile Base station (C-MB) problem, where the base station can only move to predefined points. Once the authors realize and prove the conversion from the temporal optimization problem to the location-based one, they drive a solution for the C-MB and prove that it converges to the $(1 - \epsilon)$ optimal solution. To do so, they convert the infinite number of mobile base station locations to a finite number.

First, the base station movement space is narrowed down to the smallest enclosing disk (SED) for all WSN nodes. Then, the SED is divided into subareas as follows: The possible minimum distance D_{min} and maximum distance D_{max} between

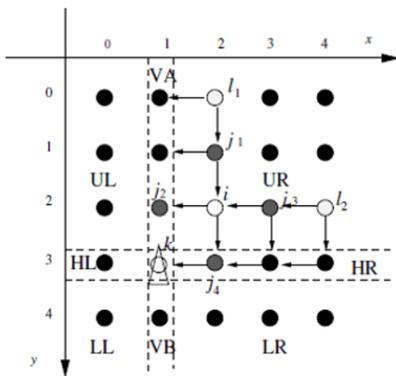


Fig. 4. Divided Subareas, and Data Flows Received and Transmitted at Node i [9]

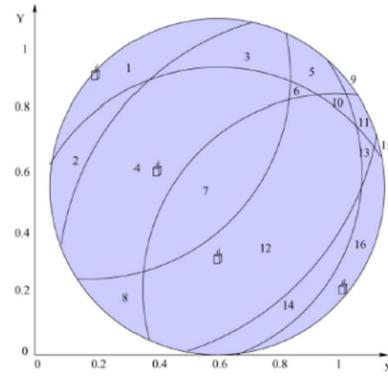


Fig. 5. WSN with Four Nodes and their Divided Sub-Areas [11].

each node i within the SED and the base station is calculated. Then, several H_i circles are drawn with a common origin at the sensor node i . The first circle's radius $D[1]$ equals to D_{min} , and the second equals to $D[1] \times (1 + \epsilon)^h$ (as shown in Fig. 5). Generally speaking, since $D[h] = (D[1] \times (1 + \epsilon)^h)$, then the energy cost can be discretized following the geometric sequence with a factor of $(1 + \epsilon)$ for each circle. The intersections between all circles of all nodes are the divided subareas. Each sub-area is presented by the fictitious cost point (FCP), which is the cost vector that embodies the upper bound of the cost of any point within this sub-area to any other node in the network. These FCP are not physical points, but they are fictionally used to prove the $(1 - \epsilon)$ -optimality of the algorithm.

Finally, the C-MB solution is applied using the FCPs in the U-MB problem. This work proved that the total sojourn duration for the base station in each sub-area is maximizing the WSN lifetime is greater than the C-MB optimal lifetime and $(1 + \epsilon)$ U-MB lifetime. This guarantees that the suggested algorithm lifetime is at least $(1 + \epsilon)$ far from the optimal lifetime obtained by the U-MB algorithm. We believe that the proposed solution will be better if it considers the optimal path for the base station mobility. Indeed, choosing the shortest path will minimize the travelling time between FCPs and the nodes' buffer size, thus maximizing the WSN lifetime.

B. Exploiting a Mobile Relay or a Special Property

Other researchers propose to use more than one mobile entity to collect data, which we refer to in our survey as the mobile relay. Moreover, some works exploit a special property (such as recharging batteries) to make the mobile entity collect data more efficiently.

In [13], the authors proposed a framework to investigate the wireless sensor network lifetime maximization based on the graph model using joint discrete sink mobility and routing problem. First, they expanded the network as two graphs: on-graph and off-graph. On-graph is when the sink is co-located at a sensor node location, while off-graph is when the sink is not. The authors proved that using predefined sink sojourn points with variant sojourn duration for the sink is a simplified version of the Maximizing the Network Lifetime (MNL) problem. Consequently, they formalized a linear program solution for the MNL problem using a single mobile sink (MNL-SMS) for the on-graph case. Then, they formulate the Primal-dual

algorithm for the set of paths between a node i and the sink, and the set of paths that go through the node i in the k^{th} sink sojourn duration after reformulating it as a path-flow. They used an extension of the Floyd–Warshall algorithm to compute the all-pairs shortest path. The resulted paths are organized into clusters, each containing paths with the same end. Then, they run a search algorithm to find the best sink sojourn duration t_k that achieves minimum energy consumption. Consequently, they generalized the primal-dual algorithm to the MNL problem using multiple mobile sinks.

Finally, they proved that using a sink layout schedule will maximize the WSN lifetime longer than maximizing the WSN lifetime by any sink with a fixed layout and referred to this problem as the TO MOVE OR NOT TO MOVE (TMNTM) decision problem. Moreover, they found that the mobile sink in the on-graph case relieves the forwarding load from the co-located sensor node and thus saves the energy consumption of that sensor node. However, the suggested algorithm still has a high computational complexity even though it is an approximated version from the Floyd–Warshall algorithm.

In [14], the authors investigated the WSN lifetime achieved using either a mobile relay or a mobile sink and their trade-offs. The mobile relay is a sensor node with unlimited energy (i.e., batteries can be replaced or recharged). First, the authors compute the WSN lifetime achieved by a static routing when the sensor nodes are all static and know the relay node location. In this case, the optimal data routing schedule is determined by sorting the static nodes' lifetime in ascending order. Then, these nodes are visited accordingly to collect data. As for the dynamic routing, where the static nodes don't know the relay node's current location to send their data, the authors proposed a linear programming model for the data routing while assuming that the relay node is co-located with a static node. The author proved that using the mobile sink will maximize the network lifetime better than the relay node. But, sometimes, it might not always be possible for the sink to be mobile as in hostile terrains. In this case, using a mobile relay will be efficient. Consequently, using a mobile relay along with a mobile sink will maximize the WSN lifetime four times compared to using a static sink. For achieving this, the mobile relay has to stay two-hop away from the sink. The authors constructed a joint mobility and routing algorithm. When the sink modifies its position, it will select a new node as an aggregation node marked to send its new position to all sensor nodes in the network. Nodes in the one-hop radius of the sink (P_1) send data to the sink directly. Those outside P_1 require to know the location of the mobile relay to send data to OM line, O being the position of the sink and M that of the mobile relay. As such, this data will be sent to the aggregation node through OM , then to the sink, which is called an Aggregation Routing Algorithm (ARA). Finally, the authors showed through simulation how using a mobile relay with a mobile sink maximizes the WSN lifetime. This approach is efficient because it applies to large-scale networks with a high density of nodes. However, since two mobile entities are used (a mobile sink and a mobile relay), building a new routing incurs more communication and computation overhead since the routing changes when their positions are changing.

Xie L et al. [15] step forward by proposing a mathematical

model to move the Wireless Charging Vehicle (WCV) which is used to both collecting data and recharging the sensor nodes. The main goal of designing their model is to minimize the entire system's energy consumption. This goal is reached by the following: 1) minimizing the energy used to move the WCV along the predefined path, which is equivalent to the maximization of the ratio of the WCV vacation time to its travelling time; and 2) minimizing the energy rate used to charge the sensor nodes, which will vary depending on each node and its distance to the WCV. Thus, to achieve these two goals, the authors proposed a mobile WCV travel path that minimizes the WCV energy while ensuring data collection from nodes and recharging all sensor nodes in a balanced manner. Beside the three main constraints in WSN they defined their own: time constraint and energy criteria constraint. The time constraint imposes that the WCV total travelling time which depends on the path distance and the velocity of WCV plus the total stopping time must be less than the WCV vacation time. As for the energy criteria constraint, it dictates to ensure that each sensor node energy which equals E_{max} at the beginning of the network ($t = 0$) must be fully charged back to E_{max} at any time of the renewable cycle before its end. Based on these constraints, the authors developed a time-dependent optimization problem. However, since the time-dependent problem is an NP-hard problem, they solved it by considering a special case of the problem that depends on the WCV's location (space-dependent) instead of the time dependency and they replaced each time variable by a space variable. Then, they proved that the time-dependent solution is equal to the space-dependent solution by showing that for any value achieved by a time-dependent feasible solution (in a time t), there is a same value achieved by the space-dependent feasible solution (in location p). That being said, the space-dependent problem has an infinite number of stop points in the WCV travelling path. So, to find the near-optimal solution, they discretize the path into a finite number of segments and assign a logical point to each segment. Then, the upper bound of each segment is determined by calculating the best case (where a node has a minimum energy consumption and is charged by the maximum energy). On the other hand, the lower bound corresponds to the worst-case (i.e., a node has a maximum energy consumption and is charged by the minimum energy). Afterwards, the gap between these two bounds is calculated; if it is not small enough, then this segment is discretized until reaching a gap within $(1 - \epsilon)$ between upper bound and lower bound. The authors proved that their algorithm finds the threshold between collecting data and recharging the sensors, and of course, this makes the WSN work continuously. A summary of all these LP models is provided in Table I

IV. ARTIFICIAL INTELLIGENCE-BASED SINK MOBILITY MODELS

In this section we will present several artificial intelligence solutions proposed to solve the energy-hole problem. Some of them are applied to the cluster-based network model while the others consider a non-cluster based architecture.

A. Non-Cluster based Models

Anwit et al. [16] proposed a variable-length chromosome genetic algorithm to find the optimal set of sink sojourn

TABLE I. SUMMARY OF LP-BASED SOLUTIONS

Ref No.	Key Concept	Advantages	Limitations
With Exploiting A Single Mobile Sink			
[8]	Distributed sink mobility algorithm. Execute locally, in parallel at each node.	It is a distributed algorithm, that means lower computation overhead	
[9]	A joint algorithm to defined sink movement, and sink sojourn duration.	Balance the energy consumption	Overhead of computation and communication is high
[11]	Continuous sink mobility solution by deriving a solution for (U-MB) problem from the (C-MB) problem solution. (Shi Y, 2012)	Near to the $(1 - \epsilon)$ optimal solution	It has a polynomial complexity
With Exploiting A Mobile Relay or A Special Property			
[13]	A primal-dual algorithm that generalized the single mobile sink solution to the multiple mobile sink problem. (Luo J, 2010)	The algorithm gives jointly mobility and routing solution for lifetime maximization, and it applies to many network topologies.	Has a computational complexity and it is not suitable for line network
[14]	Maximizing WSN lifetime using a mobile relay by determining his optimal mobility area to collect data. (Wang W, 2005)	It is appropriate to the large-scale networks that have a high density of nodes	Has a high overhead of computation and communication since there are two mobile entities.
[15]	Moving wireless charging vehicle (WCV) that used to recharge nodes and collect data from them, by developing a space-dependent algorithm.	They used WCV to recharge sensor nodes while trying to maximize the WSN lifetime.	The path must be predefined, the algorithm only found the sojourn points that maximize the WSN lifetime.

points and their locations. They applied their algorithm as follows: First, they start by generating a variable number of chromosomes from size $(0.2 \times \text{the number of sensor nodes in the WSN } (n))$ to size $(0.5 \times n)$. Each chromosome consists of a sequence of IDs numbers of random sojourn points' locations. After that, for each chromosome, they generate a population with fixed-sized. To evaluate these chromosomes, a balance between two factors is necessary. The first factor is the chromosome length which equals the number of sojourn points IDs in this chromosome. The second factor is the mobile sink path length, which is the total distance that the mobile sink cross when it moves through all these sojourn points. A high number of sojourn points may produce a longer path while there is a shorter path covering the same nodes. However, a small number of sojourn points may result in a longer path that uncovers some sensor nodes. So, to ensure the compromise between these two factors, a Fitness Function is used to filter all the unfit chromosomes that do not satisfy the threshold between making the sink path short but covering all nodes. the solution proceeds as follows:

- 1) One chromosome from two different populations is selected using the roulette wheel selection algorithm.
- 2) After that, the multi-points crossover technique is applied on these two selected chromosomes based on a predefined probability.
- 3) Then, the mutation operation is performed to the child chromosomes that were generated from the crossover operation. In the mutation operation, the gene position value to be changed is also selected using a predefined probability.

At the end of this cycle, the children's chromosomes are compared to their parents' chromosomes. If they are better, they replace them in the new population. Otherwise, they are rejected. This cycle repeats until a predefined termination condition is reached. Finally, the simulation results showed that the proposed algorithm outperforms the traditional Travel Salesman Problem (TSP) algorithm by reducing the path length and

the data collection time. However, the authors did not consider the energy consumption at each node and its impact on the path selection, which is the main goal of the sink mobility research.

In [17], a data collection strategy is proposed using a mobile sink that moves based on an ant colony optimization algorithm. First, the authors introduce a method to select the rendezvous nodes that can communicate directly with the mobile sink. These nodes are selected based on the entropy weight method of many indicators (e.g., the relative residual energy and density of the nodes) to select the optimal access path for the mobile sink. This optimal solution should minimize the WSN energy consumption and tolerate some delay to compensate for energy consumption. The simulation of the proposed algorithm showed the energy consumption is balanced between all sensor nodes. Nevertheless, the main limitation of this research is that the solution focuses on minimizing the whole network residual energy amount instead of maximizing the network lifetime.

Zhu et al. [18] proposed a WSN scheme inspired by the honey-comb structure to collect data from the WSN sensor nodes. First, they assumed that the first position of the sink is the WSN's centre point. Then, they divide the network into a number of hexagons, each hexagon is called a partition. Each partition must contain at least one sensor node and its side length (a) must be bigger than or equal to the power two of the sensor node's sensing radius (r^2). Besides, they established the Cartesian coordinate system in the network with an origin O located at the WSN's centre point. After the network partition, they gave each hexagon a partition ID and a direction value. Like the Cartesian coordinate, the partition ID is represented by two integers (m, n) , where m is increasing along the x -axis and n is increasing along the y -axis progressively. It is worth noting that n increases by two for the vertical partitions and by 1 for the non-vertical partitions. As for the direction, each partition's direction value is determined based on the sink position. The sink position is at the Cartesian coordinate centre point $(0,0)$ and its direction value is 0. As the suggested scheme

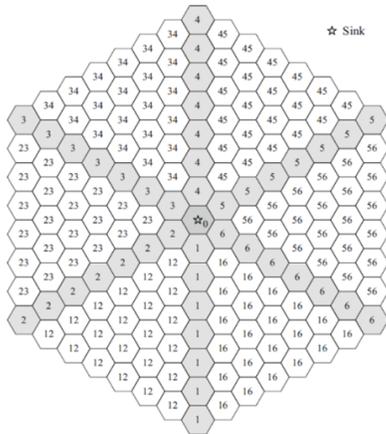


Fig. 6. Direction Value for Each Partition [18]

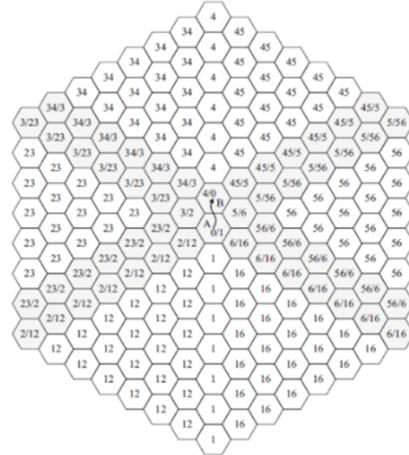


Fig. 7. Updating Directional Value for Partitions in Sink Vertical Movement Case. [18]

consists of many hexagons, each edge of the sink's hexagon is represented by a number n from 1 to 6. When the hexagon partition is located on the (n^{th}) side direction, the direction value of these partitions is equal to the (n^{th}) number. The remaining partitions have a combined direction value, i.e. the values of the surrounding partitions (e.g. the partition between partitions 1 and 2 has the value 12 as shown in Fig. 6).

Regarding the sensor nodes, they will be assigned a combined addressing values containing (i) partition ID, (ii) the direction value, and (iii) the node ID which is a unique number used to distinguish sensor nodes. Once every sensor is assigned to a given partition, then there are three aspects of this data-gathering scheme: 1) data forwarding between partitions, 2) sink moving strategy and 3) updating the direction value. For the data forwarding, the node will forward its data to the sink partition through its direction value. For the sink moving strategy, first of all, the sink has a counter to count the number of received packets; if it reaches a threshold, then the sink will move and change its location. There are three strategies for sink movement:

- 1) The random movement (RM), where the sink selects one of its partition's neighbours randomly.
- 2) The data volume-based greedy movement (DGM), where the sink has counters on each hexagon side, used to count the number of the received data packets, then the sink chooses the side with the highest counter to move to.
- 3) The energy-based greedy movement (EGM), where the sink employs an agent node for each hexagon side to calculate the average residual energy for sensor nodes on that side, then the sink will choose to move to the side with the highest residual energy.
- 4) For the direction value update, the sink moves either vertically (to sides numbered by 1 or 4) or non vertically (to sides numbered by 2, 3, 5, or 6). When the sink moves, it discloses its new location to some partitions. Indeed, there is no need to broadcast the new position to all sensors since the many partitions will still locate at the same partition regarding the sink position, thus achieving energy efficiency (see Fig. 7).

Finally, the author simulated their three proposed schemes using MATLAB, then they compared them based on 1) sensor nodes density, 2) ratio of the source nodes in WSN, and 3) the hexagon side length. The main performance criteria are the average energy consumption, the maintenance cost, packet collected, and packet loss rate. They found that their HSDG scheme with DGM for the sink (HSDG-DGM) is outperforming other data gathering schemes HSDG-EGM and HSDG-RM in the energy-efficiency with low data loss. As an advantage, this algorithm is energy efficient since the broadcast and update of the sink's new location is limited to a given direction partition. Moreover, it avoids the routing void. Indeed, when there is no sensor node to forward the data to in a partition, the forwarding scheme states that the data packet is transmitted to another neighbour partition.

Gharaei et al. [19] proposed Energy-Efficient Mobile sink Sojourn Location Optimization (EMSLO) algorithm. Their algorithm consists of two phases: 1) Evaluating sensor nodes residual energy, and 2) Optimizing sink sojourn points. First, to evaluate the sensor nodes in the network, the remaining alive time for each sensor is calculated by dividing the residual energy by the energy consumption rate. After that, the sensor with the minimum lifetime is chosen to be the Critical Node (CN). Next, all nodes are given a weight calculated by dividing the CN lifetime by its lifetime. Accordingly, the CN is the start point of the sink sojourn location, to find the next point, the genetic algorithm is used to select the sensor nodes with the minimum cost. This cost is calculated by a function based on two factors: 1) The variance of residual energy of sensor nodes. 2) each sensor weight that has been calculated before depending on CN lifetime. Finally, the authors simulate their scheme and found that their model decreases the variance of sensor residual energy, which means a good distribution of energy consumption. Moreover, they compare their model to EPMS [20] and DCHS [21] in terms of WSN lifetime, and found that their model EMSLO maximizes the WSN lifetime more than others. Moreover, they compare their algorithm to CMS2TO [22] and CM2SV2 [23] to prove that optimizing the sink sojourn location is better than optimizing the sojourn duration or the speed of the sink, and the result emphasizes that.

B. Cluster based Models

Lu et al. [24] and Jong et al. [25] both inspired from bees and queen honeybee behaviour respectively. First, in [24] they proposed a path optimization strategy for the mobile sink in WSNs using an artificial bee colony algorithm. They improve the data gathering ability by obtaining the optimal trajectory design of the mobile sink. First of all, they cluster the nodes in the grid and select some nodes as cluster heads by the traditional method LEACH [26]. These cluster heads are considered as the rendezvous points of the mobile sink. Then, they transform the WSN energy consumption problem into a minimization of the total hops between all rendezvous points and the sensor nodes by establishing the constraints' criterion that makes the WSN lifetime maximized. Moreover, they optimize the artificial bee colony (ABC) algorithm to solve this problem. As the ABC algorithm, where employed bees are randomly selected to search for the initial food sources, the authors proposed a formula to select the initial rendezvous points. After selecting the initial individual, they compare their fitnesses to a newly generated individual to determine which one is better. If the new individual is better, then they replace the old with it, if not then increase the cumulative factor by one. This cumulative factor is used to indicate the quality of this individual, the more quality they have, the higher its probability to be selected. Then the food source or the rendezvous points will be selected using the roulette method based on their probabilities. Besides, the authors used the Cauchy mutation detection operator based on the current solution. By using this operator, the feasible solution diversity is increased, and the randomness of the solution is avoided. Moreover, the convergence to the optimal solution is achieved, which will increase the accuracy of the solution.

Jong et al. [25] proposed a scheme to maximize the WSN lifetime inspired by the migration process of the Queen called a QHBM algorithm. First of all, based on the nature of the Queen migration such as travel for food, depending on scout bees to look for new food sources or places, follow the scout bees with high sign (excitement), stop to rest at several points, and repeat all of these till they decide where to build its hive. By applying this scenario in the network, let the Queen be a sink. For the sensor nodes, they are clustered randomly. Each cluster has some CH sensor nodes selected randomly too. These CH nodes are responsible for forwarding the cluster sensor nodes data to the sink, if their residual energy becomes below a threshold, they will send an alert to the sink to start its journey and leave this cluster, so this cluster could replace the weak CH nodes and assign the nodes with the highest residual energy as the new CH nodes. In the proposed algorithm, these CH nodes will take place of bee scouts, they will lead the sink to come over. To mimic the honeybee migration, by knowing that the Queen honeybee travel to one of the 8 cardinal directions poles, and between every two poles, there is a sector containing bees scouts. Moreover, let the residual energy of the CH nodes become the sign (excitement) of the scout bees. Consequently, the proposed algorithm steps will be as follows:

- 1) Set the variables: sink initial position, sink communication radius, the confidence factor.
- 2) Assign the scouts from the CH nodes.
- 3) Group the CH nodes located in the same sector.
- 4) The sink calculates the average residual energy of the

CH nodes in each sector.

- 5) Then the sink calculates each sector probability. After that, the sink selects the sector that has the highest probability, which means that this sector has the CH nodes with the highest residual energy.
- 6) After selecting the sector, the sink compares the probability of the two adjacent sectors of the selecting one, then selects the pole in the edge of the sector with the highest probability as a destination.
- 7) Finally, the sink start its journey after receives an alert from the CH nodes.

The authors proved that their proposed algorithm is maximizing the WSN lifetime compared to a static sink since the re-position of the sink is alleviating the energy-depletion on the same nodes. Moreover, QHBM algorithm is alleviating the routing overhead since the sink only communicates with the CH nodes, so only these nodes require the update of the sink position. Additionally, the QHBM provides a balanced energy consumption since there is a scanning phase, where the sink looks for the CH nodes that have the highest residual energy to travel to. As a limitation, the authors did not find the optimal value of the sink communication radius, and the confidence factor is still calculated randomly.

Kumar et al. [27] proposed an Energy-Aware sink Mobility model (EASM). First of all, they proposed to logically divide the grid into a number of sub-grids, each sub-grid has an ID number, and this ID number is embedded also in each sensor node ID that belongs to it. After that, to determine the movement direction of the sink, the sink first calculates the average residual energy of its one-hop neighbors at each sub-grid. Then, it categories these sub-grids into 4 levels of energy called: Adequate Level, Operation Level, Warning Level and Danger Level. If the energy level of the sojourn grid reaches the warning or danger level, then the sink has to move by calculating each sub-grid's average residual energy then selecting the grid with the maximum residual energy level as a destination. After the sink reaches the new sub-grid, it applies the Breadth-First Search (BFS) technique to generate a spanning tree that connects all sensors in that sub-grid to the sink with the shortest path. Finally, the authors simulate their algorithm and compare it to a static sink and Energy-Aware Sink Relocated model in terms of the number of alive sensors, average residual energy, and amount of collected data. They found that their algorithm EASM is outperforming the others.

Wang et al. [28] proposed a path planning model for the mobile sink to minimize the energy consumption of the sensor nodes and make the sink travel path as short as possible. Namely, to create this path, a number of Rendezvous Points (RP) were selected after four phases: 1) tree formation, 2) RP candidate, 3) RP selecting and 4) Finding the shortest path. The tree formation is done simply by applying the spanning tree algorithm. While for the RP candidate, they follow the bottom-up manner and check the buffer remaining space for each sensor while considering the amount of data that comes from its child sensors. If a node with a full buffer capacity is found, then this node will candidate to be one of the final RP. After that, each candidate RP with its sub-tree will be evaluated by Depth-First Search (DFS) based on hop count and distance. Finally, by connecting these RP and finding the TSP, the travel path of the mobile sink is created. Moreover,

the authors proposed an enhancement of their model to further reduce the travel path length. Their enhancement is done by finding better RP by replacing it or combining two of them depending on some calculation of its RP and RP-1 and RP+1.

Finally, the authors evaluate their model in terms of the dropped packet, energy consumption, computational time and buffer exploitation. They proved that their model exploits the RP buffer better than other models (CB-E [29] and WRP [30]). Consequently, the number of dropped packets is minimized. Moreover, they proved that their model is energy and computationally efficient.

Table II summarizes the AI-based solutions. It provides the key concept, advantages and disadvantages of each work.

V. OTHER SINK MOBILITY MODELS

Some of the proposed models are dependent on the sinks statistical and operational decisions made for moving. Shah et al. [31] proposed a model using Mobile Ubiquitous LAN Extensions (MULEs). MULEs also are not suitable for a real-time data system. First of all, the authors consider that the network consists of (i) several Access Points, which are uniformly distributed, (ii) several sensor nodes, randomly distributed, and (iii) several MULEs. Additionally, they assumed that:

- The MULEs move with 0.25 probability to any of the four main directions of the grid (North, South, East, West) with every global clock tick.
- The MULEs can't communicate or exchange data with other sensor nodes or access points unless they are co-located at the same grid points.
- Due to the limitation in MULEs' and sensor nodes' buffer size, any amount of data transferred to them will be dropped, if it is full.

They relied on Markov chains and its transition probabilities to estimate the average values of the inter-arrival time of the MULE to a sensor, the number of steps that the MULE takes to complete one round, and the average number of data amount that MULE picks up and transfer to the AP. After that, to measure the performance, they consider the following criteria: Data Success Rate and Buffer Size (keep it as small as possible) From the performance result, the authors found that:

- The sensor buffer size is increasing linearly with the grid size.
- Data latency at sensor nodes is increasing linearly with the sensor nodes' buffer size.
- MULEs buffer size is increasing with the square of the grid size.
- Data latency to reach the access point is increasing linearly with the grid size.

As a result, if the grid size increases, it is required to use multiple MULEs and multiple access points, to achieve a high success data rate while keeping an appropriate buffer size. Indeed, using a large buffer size will lead to energy consumption, which is against the goal of maximizing the

WSN lifetime. As benefits of using MULEs, it can increase scalability, flexibility, and robustness for the WSN. But it may have some drawbacks because of its unpredictable and constrained movement. Indeed, sometimes the MULE is unable to reach the sensor nodes due to the change in terrain that causes limitations in the MULEs movement.

Gosh et al. [32] proposed a new sink mobility approach based on the sensors request. First of all, they assume that the sink has a limited queue with size n . When a sensor node's buffer becomes near to be full, it sends a request to the sink to move and collect data directly through a single hop. On the sink side, the sink save these request on its queue and serve them based on two scheduling schemes: 1) First Come First Serve (FCFS), 2) Nearest Job Next (NHN). The main contribution of the authors is joining new schemes to the previous two. Thus, additionally there will be 3) Earliest Deadline First-First Come First Serve (EDF-FCFS), and 4) Earliest Deadline First-Nearest Job Next (EDF-NHN).

First, it is important to point out that each request has to be served before its deadline. Moreover, each request has a service time, which is the time taken to completely transmit the data from the sensor's buffer to the sink. There is also a movement time, which is the time taken by the sink when it moves between the sensors based on the serving queue order. Due to the different sensor locations and the nodes' geometric distribution, the order of serving the sensors' requests is a critical thing, since the sink moving time will be considered. Accordingly, the authors first apply the first two schemes: FCFS and NHN, if all requests in the queue will be served before their deadlines, then this queue order is good and there is no need to change it. Else, if there is a request i that will miss its deadline, the two newly proposed approaches will be applied: EDF-FCFS and EDF-NHN. According to these approaches, if there is a request i that will miss its deadline, then this request will reverse the order with the request $(i-1)$, if it is still, then reverses it with $(i-2)$, if it is still, then this request will not be served and dropped from the queue. Finally, the author evaluated the four schemes and compared it based on the number of not-served request and the response time. They proved that their adapted schemes (EDF-FCFS and EDF-NHN) have outperformed the originals (FCFS and NHN) with less number of not-served requests and faster responses. But, still, these schemes are not suitable for large scale networks.

Table III summarizes the previously-mentioned solutions.

VI. DISCUSSION

We compare in Table. IV the research works that were discussed earlier are based on the following criteria:

- *Network Topology*: many types of networks have been investigated including (i) grid (regular) network, (ii) random network, and (iii) scale-free network. The network topology of the reviewed papers varies between grid and random. Some of the random typologies are clustered by (i.e. See SectionIV-B)
- *Sink Sojourn Points*: sojourn points are points where the sink stops to collect data from the sensor nodes. Most of the reviewed papers under LP models are using predefined sojourn points while the AI-based

TABLE II. SUMMARY OF AI-BASED SOLUTIONS

Ref No.	Key Concept	Advantages	Limitations
Non-Cluster Based Models			
[16]	A genetic algorithm to find the optimal number of sink sojourn points and their locations	It is used to find the shortest path	The energy consumption at each node is not considered, which of course impact on the path selection.
[17]	A data collection strategy Using ant colony optimization algorithm to move the sink.	It is used to find the best path between the sojourn points.	It minimizes the whole network residual energy consumption, instead of maximizing the network lifetime.
[18]	Sink mobile through a WSN scheme inspired by the honey-comb for data collection	Energy efficient, and avoiding the routing avoid	
[19]	Use genetic algorithm to select sink sojourn points after calculating the cost for each sensor nodes	Easy to implement, applicable for high density WSN.	
Cluster Based Models			
[24]	Moving the sink using Bee colony optimization algorithm.	Avoid the randomness of the solution Using the Cauchy mutation detection	
[25]	QHBM algorithm inspired by the migration process of the Queen Honey Bee	Due to the scanning phase in the algorithm, it provides a balancing energy consumption	Many variables need to be restricted, such as the confidence factor which is calculated randomly.
[27]	First, divided the grid to sub grids, then calculate the average residual energy and determine the energy level for each sub-grid. After that if the sojourn sub-grid energy level decreased to 4th, the sink mobile to the sub grid with the highest energy level.	Efficient for high density WSN.	
[28]	Applying the spanning tree, then go bottom-up and check the buffer to find sojourn points.	It ensures transmission and exploits the whole amount of buffers.	

TABLE III. SUMMARY OF OTHER SINK MOBILITY MODELS SOLUTIONS

Ref No.	Key Concept	Advantages	Limitations
[31]	MULEs mobility routing, Based on the Markov chains transition probabilities.	Easy to implement	The MULE is unpredictable and physically moved so it is affected by the changes in its environment.
[32]	Sink Mobility based on two sachems of serving sensor request EDF-FCFS and EDF-NJN.	Easy to implement	Many packet will be loose because the sink has a limited queue size, and the request has a deadline. Hence, it is not scalable.

solutions define the sojourn points randomly (*e.g.*, [16]), actively (*e.g.*, [18]), or based on sensor query (*e.g.*, [32]).

- *Sink Sojourn Duration*: it refers to the time the sink takes at a sojourn point to complete collecting data from the sensor nodes. The WSN lifetime is actually the sum of all sink sojourn duration. So it is important to find the trade-off between maximizing the sojourn duration while taking into account the delay-tolerance time.
- *Buffer size*: it is the size of the sensors' temporary storage that is used to store data until the sink comes. All reviewed papers assume that the buffer size is infinite, though it is an unrealistic assumption.
- *Overhearing*: in WSN, the nodes are always listening to the traffic in their transmission range, hence consume an additional amount of energy. However, none of the reviewed papers has considered the overhearing while it is an important source of energy consumption that should be taken into account in WSN.

VII. FUTURE RESEARCH DIRECTIONS

We provide in the following some insights on possible future research directions to pave the way towards building more realistic and robust solutions for the energy hole problem.

A. Considering Additional Constraints

Based on the aforementioned comparative analysis, we deduce that two constraints are always considered in all WSN-lifetime-maximization solutions namely (i) the energy constraint and (ii) the flow conservation law. Consequently, it is highly recommended to consider more constraints that impact the WSN lifetime such as:

- *Buffer Size*: The buffer size is an important criterion as it constrains the sensor capacity to store received and generated data packets which has an impact on the sink sojourn duration at each sojourn point. Besides, a higher buffer size leads to a higher energy consumption. So, finding the optimal sojourn duration that takes into account the size of the buffers that avoid data rejection is one of the main goals

TABLE IV. COMPARISON OF SINK MOBILITY MODELS

Taxonomy	Reference No.	Network Topology	Sojourn Points	Sojourn Duration	Buffer Size	Overhearing
LP	[8]	Random	✓	X	X	X
	[9]	Grid	✓	✓	X	X
	[11]	Random	✓	✓	X	X
	[13]	Grid	✓	✓	X	X
	[14]	Random	✓	✓	X	X
	[15]	Random	✓	X	X	X
AI	[16]	Random	✓	X	X	X
	[17]	Random	✓	X	X	X
	[18]	Random	✓	✓	X	X
	[19]	Random	✓	X	X	X
	[24]	Grid	✓	X	X	X
	[25]	Random	✓	✓	X	X
	[27]	Random	✓	✓	X	X
	[28]	Random	✓	X	X	X
OM	[31]	Grid	✓	X	X	X
	[32]	Random	✓	✓	X	X

for WSN-lifetime-maximization [33] [31]. However, this constraint is generally ignored when solving the energy hole problem.

- *Overhearing Energy Consumption:* Since we deal with wireless sensor networks, ignoring the overhearing is unrealistic and gives flawed results for the WSN lifetime as it inevitably consumes energy. However, none of the reviewed papers considers overhearing. Thus, it is important to devise new models that take it into account when searching for the optimal sojourn duration for the sink using techniques that reduce the overhearing as in [34].
- *Link Capacity:* Similarly to the buffer size, the link capacity constrains the number of received and transmitted data packets, which consequently impact the energy consumption since it is calculated per data unit. As such, new models that consider this constraint are required in the field.

B. Reducing Latency

Even though the reviewed models are applicable for delay-tolerant WSN, reducing latency is important to avoid the data loss due to the buffer size limitation. Many approaches have been suggested to reduce the latency by using multiple mobile sinks that cover the WSN simultaneously, which allows to collect data faster. However, this solution has a higher cost which is undesirable in WSN [35].

Another alternative is to use a powerful mobile sink that can move between sensor nodes at a higher velocity. Similarly to the previous approach, this will be more efficient than a single sink but would incur higher costs. Meanwhile, as the WSN is usually used in harsh environments, the moving in a high speed can be difficult to achieve.

Consequently, more research is required to find the trade-off between the efficiency and the total cost of the proposed solution while taking into account the geographical characteristic of the environment and the buffer size constraint.

C. Utilizing Additional Energy Resources

One of the approaches to prolong the WSN lifetime is to recharge sensors' batteries. Even in a harsh environment,

there are many ways to recharge sensors such as using the solar power, microwave energy transfer, radio-frequency energy transfer, and wireless charging stations where the mobile sink could be used as a recharging station [15]. Joint data gathering using mobile sink and sensors' battery recharging is a promising solution to optimize the WSN lifetime [36] [37][38][39]. Even though lots of these approaches had been proposed, few works combine both techniques. A possible future research direction would be to investigate the strengths of both techniques (*i.e.*, maximizing the WSN lifetime and energy recharging) to design a promising WSN lifetime optimization solution.

VIII. CONCLUSION

In this survey, numerous sink mobility models were discussed in the literature review to mitigate the WSN energy hole problem and maximize the WSN lifetime. These sink mobility models were classified into three categories: Linear programming-based solutions (LP-based), Artificial Intelligence-based solutions (AI-based), and other computation models. Regarding the LP-based models, four solutions used only one mobile sink, while the rest used either an additional entity or a special property (*e.g.*, recharging sensor nodes). As for the AI-based models, they have two categories: cluster based models (*i.e.*, sensors are organized into clusters, where each cluster has a cluster head), and a non-cluster based models. Finally, the last section discusses additional solutions that neither used LP nor AI-based.

All the solutions are discussed in detail and compared according to a set of criteria. The comparative analysis shows that most of the research ignored important constraints that fully impact WSN lifetime, namely, limited buffer size, overhearing, and link capacity. Finally, future research directions were suggested to develop a more comprehensive WSN lifetime maximization using a mobile sink.

REFERENCES

- [1] G. Xu, W. Shen, and X. Wang, "Applications of Wireless Sensor Networks in Marine Environment Monitoring: A Survey," *Sensors*, vol. 14, no. 9, pp. 16 932–16 954, Sep. 2014, number: 9 Publisher: Multidisciplinary Digital Publishing Institute. [Online]. Available: <https://www.mdpi.com/1424-8220/14/9/16932>

- [2] A. A. Abbasi and M. Younis, "A survey on clustering algorithms for wireless sensor networks," *Computer Communications*, vol. 30, no. 14, pp. 2826–2841, Oct. 2007. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S0140366407002162>
- [3] Q. Dong and W. Dargie, "A Survey on Mobility and Mobility-Aware MAC Protocols in Wireless Sensor Networks," *IEEE Communications Surveys Tutorials*, vol. 15, no. 1, pp. 88–100, 2013, conference Name: IEEE Communications Surveys Tutorials.
- [4] J. Li and P. Mohapatra, "Analytical modeling and mitigation techniques for the energy hole problem in sensor networks," *Pervasive and Mobile Computing*, vol. 3, no. 3, pp. 233–254, Jun. 2007. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S1574119206000630>
- [5] S. Yu, B. Zhang, C. Li, and H. T. Mouftah, "Routing protocols for wireless sensor networks with mobile sinks: a survey," *IEEE Communications Magazine*, vol. 52, no. 7, pp. 150–157, Jul. 2014, conference Name: IEEE Communications Magazine.
- [6] A. Jarry, P. Leone, O. Powell, and J. Rolim, "An Optimal Data Propagation Algorithm for Maximizing the Lifespan of Sensor Networks," in *Distributed Computing in Sensor Systems*, ser. Lecture Notes in Computer Science, P. B. Gibbons, T. Abdelzaher, J. Aspnes, and R. Rao, Eds. Berlin, Heidelberg: Springer, 2006, pp. 405–421.
- [7] "A new energy efficient and fault-tolerant protocol for data propagation in Smart Dust networks using varying transmission range - IEEE Conference Publication." [Online]. Available: <https://ieeexplore.ieee.org/abstract/document/1299464/>
- [8] Y. Yun, Y. Xia, B. Behdani, and J. C. Smith, "Distributed Algorithm for Lifetime Maximization in a Delay-Tolerant Wireless Sensor Network with a Mobile Sink," *IEEE Transactions on Mobile Computing*, vol. 12, no. 10, pp. 1920–1930, Oct. 2013, conference Name: IEEE Transactions on Mobile Computing.
- [9] Z. M. Wang, S. Basagni, E. Melachrinoudis, and C. Petrioli, "Exploiting Sink Mobility for Maximizing Sensor Networks Lifetime," in *Proceedings of the 38th Annual Hawaii International Conference on System Sciences*, Jan. 2005, pp. 287a–287a, iSSN: 1530-1605.
- [10] "Lindo Systems." <http://www.lindo.com>, 2008, [Online; accessed 19-July-2008].
- [11] "Some Fundamental Results on Base Station Movement Problem for Wireless Sensor Networks - IEEE Journals & Magazine." [Online]. Available: <https://ieeexplore.ieee.org/abstract/document/6068265/>
- [12] Y. Shi and Y. T. Hou, "Theoretical Results on Base Station Movement Problem for Sensor Network," in *IEEE INFOCOM 2008 - The 27th Conference on Computer Communications*, Apr. 2008, pp. 1–5, iSSN: 0743-166X.
- [13] J. Luo and J. Hubaux, "Joint Sink Mobility and Routing to Maximize the Lifetime of Wireless Sensor Networks: The Case of Constrained Mobility," *IEEE/ACM Transactions on Networking*, vol. 18, no. 3, pp. 871–884, Jun. 2010, conference Name: IEEE/ACM Transactions on Networking.
- [14] "Using mobile relays to prolong the lifetime of wireless sensor networks | Proceedings of the 11th annual international conference on Mobile computing and networking." [Online]. Available: <https://dl.acm.org/doi/abs/10.1145/1080829.1080858>
- [15] "A Mobile Platform for Wireless Charging and Data Collection in Sensor Networks - IEEE Journals & Magazine." [Online]. Available: <https://ieeexplore.ieee.org/document/7008490>
- [16] R. Anwit and P. K. Jana, "A Variable Length Genetic Algorithm approach to Optimize Data Collection using Mobile Sink in Wireless Sensor Networks," in *2018 5th International Conference on Signal Processing and Integrated Networks (SPIN)*, Feb. 2018, pp. 73–77.
- [17] H. Zhang, Z. Li, W. Shu, and J. Chou, "Ant colony optimization algorithm based on mobile sink data collection in industrial wireless sensor networks," *EURASIP Journal on Wireless Communications and Networking*, vol. 2019, no. 1, p. 152, Jun. 2019. [Online]. Available: <https://doi.org/10.1186/s13638-019-1472-7>
- [18] C. Zhu, G. Han, and H. Zhang, "A honeycomb structure based data gathering scheme with a mobile sink for wireless sensor networks," *Peer-to-Peer Networking and Applications*, vol. 10, no. 3, pp. 484–499, May 2017. [Online]. Available: <https://doi.org/10.1007/s12083-016-0496-6>
- [19] N. Gharaei, S. J. Malebary, K. Abu Bakar, S. Z. Mohd Hashim, S. Ashfaq Butt, and G. Sahar, "Energy-Efficient Mobile-Sink Sojourn Location Optimization Scheme for Consumer Home Networks," *IEEE Access*, vol. 7, pp. 112079–112086, 2019, conference Name: IEEE Access.
- [20] J. Wang, Y. Cao, B. Li, H.-j. Kim, and S. Lee, "Particle swarm optimization based clustering algorithm with mobile sink for WSNs," *Future Generation Computer Systems*, vol. 76, pp. 452–457, Nov. 2017. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0167739X16302540>
- [21] O. M. Alia, "Dynamic relocation of mobile base station in wireless sensor networks using a cluster-based harmony search algorithm," *Information Sciences*, vol. 385–386, pp. 76–95, Apr. 2017. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0020025516322824>
- [22] N. Gharaei, K. Abu Bakar, S. Z. Mohd Hashim, A. Hosseingholi Pourasl, and S. Ashfaq Butt, "Collaborative Mobile Sink Sojourn Time Optimization Scheme for Cluster-Based Wireless Sensor Networks," *IEEE Sensors Journal*, vol. 18, no. 16, pp. 6669–6676, Aug. 2018, conference Name: IEEE Sensors Journal.
- [23] N. Gharaei, K. Abu Bakar, S. Z. Mohd Hashim, A. Hosseingholi Pourasl, M. Siraj, and T. Darwish, "An Energy-Efficient Mobile Sink-Based Unequal Clustering Mechanism for WSNs," *Sensors*, vol. 17, no. 8, p. 1858, Aug. 2017, number: 8 Publisher: Multidisciplinary Digital Publishing Institute. [Online]. Available: <https://www.mdpi.com/1424-8220/17/8/1858>
- [24] Y. Lu, N. Sun, and X. Pan, "Mobile Sink-Based Path Optimization Strategy in Wireless Sensor Networks Using Artificial Bee Colony Algorithm," *IEEE Access*, vol. 7, pp. 11 668–11 678, 2019, conference Name: IEEE Access.
- [25] G.-J. Jong, Aripriharta, Hendrick, and G.-J. Horng, "A Novel Queen Honey Bee Migration (QHBM) Algorithm for Sink Repositioning in Wireless Sensor Network," *Wireless Personal Communications*, vol. 95, no. 3, pp. 3209–3232, Aug. 2017. [Online]. Available: <https://doi.org/10.1007/s11277-017-3991-z>
- [26] W. R. Heinzelman, A. Chandrakasan, and H. Balakrishnan, "Energy-efficient communication protocol for wireless microsensor networks," in *Proceedings of the 33rd Annual Hawaii International Conference on System Sciences*, Jan. 2000, pp. 10 pp. vol.2–.
- [27] H. Kumar, N. E., and D. K. S.M., "EASM: Energy-Aware Sink Mobility Algorithm to Prolong Network Lifetime in WSN," in *2019 IEEE 16th India Council International Conference (INDICON)*, Dec. 2019, pp. 1–4, iSSN: 2325-9418.
- [28] Y.-C. Wang and K.-C. Chen, "Efficient Path Planning for a Mobile Sink to Reliably Gather Data from Sensors with Diverse Sensing Rates and Limited Buffers," *IEEE Transactions on Mobile Computing*, vol. 18, no. 7, pp. 1527–1540, Jul. 2019, conference Name: IEEE Transactions on Mobile Computing.
- [29] K. Almi'ani, A. Viglas, and L. Libman, "Energy-efficient data gathering with tour length-constrained mobile elements in wireless sensor networks," in *IEEE Local Computer Network Conference*, Oct. 2010, pp. 582–589, iSSN: 0742-1303.
- [30] H. Salarian, K.-W. Chin, and F. Naghdy, "An Energy-Efficient Mobile-Sink Path Selection Strategy for Wireless Sensor Networks," *IEEE Transactions on Vehicular Technology*, vol. 63, no. 5, pp. 2407–2419, Jun. 2014, conference Name: IEEE Transactions on Vehicular Technology.
- [31] R. C. Shah, S. Roy, S. Jain, and W. Brunette, "Data MULEs: modeling and analysis of a three-tier architecture for sparse sensor networks," *Ad Hoc Networks*, vol. 1, no. 2, pp. 215–233, Sep. 2003. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S1570870503000039>
- [32] N. Ghosh and I. Banerjee, "Application of mobile sink in wireless sensor networks," in *2018 10th International Conference on Communication Systems Networks (COMSNETS)*, Jan. 2018, pp. 507–509, iSSN: 2155-2509.
- [33] T. Rault, A. Bouabdallah, and Y. Challal, "WSN lifetime optimization through controlled sink mobility and packet buffering," in *Global Information Infrastructure Symposium - GIIS 2013*, Oct. 2013, pp. 1–6, iSSN: 2150-329X.
- [34] A. Pal and A. Nasipuri, "DRCS: A Distributed routing and channel selection scheme for multi-channel wireless sensor networks," in *PER-COM Workshop*, 2013.

- [35] A. W. Khan, A. H. Abdullah, M. H. Anisi, and J. I. Bangash, "A Comprehensive Study of Data Collection Schemes Using Mobile Sinks in Wireless Sensor Networks," *Sensors*, vol. 14, no. 2, pp. 2510–2548, Feb. 2014, number: 2 Publisher: Multidisciplinary Digital Publishing Institute. [Online]. Available: <https://www.mdpi.com/1424-8220/14/2/2510>
- [36] M. Zhang and W. Cai, "Data Collecting and Energy Charging Oriented Mobile Path Design for Rechargeable Wireless Sensor Networks," *Journal of Sensors*, vol. 2022, p. e5004507, Apr. 2022, publisher: Hindawi. [Online]. Available: <https://www.hindawi.com/journals/js/2022/5004507/>
- [37] Q. Wu, P. Sun, and A. Boukerche, "A Novel Joint Data Gathering and Wireless Charging Scheme for Sustainable Wireless Sensor Networks," in *ICC 2020 - 2020 IEEE International Conference on Communications (ICC)*, Jun. 2020, pp. 1–6, iSSN: 1938-1883.
- [38] A. Boukerche, Q. Wu, and P. Sun, "A Novel Joint Optimization Method Based on Mobile Data Collection for Wireless Rechargeable Sensor Networks," *IEEE Transactions on Green Communications and Networking*, vol. 5, no. 3, pp. 1610–1622, Sep. 2021, conference Name: IEEE Transactions on Green Communications and Networking.
- [39] M. Angurala, M. Bala, and S. S. Bamber, "Wireless battery recharging through UAV in wireless sensor networks," *Egyptian Informatics Journal*, vol. 23, no. 1, pp. 21–31, Mar. 2022. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S1110866521000281>

End-to-End Car Make and Model Classification using Compound Scaling and Transfer Learning

Omar BOURJA¹, Abdelilah MAACH², Zineb ZANNOUTI³, Hatim DERROUZ⁴,
Hamza MEKHZOUM⁵, Hamd AIT ABDELALI⁶, Rachid OULAD HAJ THAMI⁷, François BOURZEIX⁸
RIME Department, Mohammadia School of Engineers, Mohammed V University in Rabat, 10100, Morocco^{1,2}
Department of Electronics and Informatics (ETRO), Vrije Universiteit Brussel, Brussels⁵
Embedded Systems and Artificial Intelligence Department, MAScIR, 10100, Morocco^{1,3,4,6,8}
IRDA team, ADMIR Laboratory, Rabat IT center, ENSIAS, Mohammed V University, Rabat 10100, Morocco^{3,4,7}
Corresponding Author: Omar BOURJA

Abstract—Recently, Morocco has started to invest in IoT systems to transform our cities into smart cities that will promote economic growth and make life easier for citizens. One of the most vital addition is intelligent transportation systems which represent the foundation of a smart city. However, the problem often faced in such systems is the recognition of entities, in our case, car and model makes. This paper proposes an approach that identifies makes and models for cars using transfer learning and a workflow that first enhances image quality and quantity by data augmentation and then feeds the newly generated data into a deep learning model with a scaling feature—that is, compound scaling. In addition, we developed a web interface using the FLASK API to make real-time predictions. The results obtained were 80% accuracy, fine-tuning it to an accuracy rate of 90% on unseen data. Our framework is trained on the commonly used Stanford Cars dataset.

Keywords—Vehicles classification; deep learning; compound scaling; transfer learning; IoT

I. INTRODUCTION

Intelligent Transportation Systems (ITS) represent a combination of advanced information and communication technologies. They are used in transportation and traffic management systems to enhance road transportation networks' safety, efficiency, sustainability, reduce congestion, and improve the driving experience. The performance of road networks can be monitored and adjusted in real-time. Video surveillance systems have become so used in ITS. With the massive adoption of high-definition cameras, advanced analytics, and AI, surveillance systems are faced with increased workloads and are no longer just for security purposes. An intelligent transportation system should include the minimum requirements for managing traffic, including vehicle detection [1], vehicle tracking [2], [3], and vehicle type classification [4], [5]. Because of traffic jams, lack of vehicle parking spots, and pollution, traffic control has always been a problem in the urban areas of Morocco [6]. Due to such events, traffic monitoring is crucial for collecting statistical data to design better and plan transportation infrastructure, other functionalities that can be integrated in an intelligent transport system is inter-vehicle distance estimation [7], [8]. Vehicle classification can solve numerous problems and help for a better traffic organization. Motivated by this fact, we have developed a framework to classify models and car makes in real-time to solve practical use case scenarios in ITS. In general,

identifying car make and model has not been an easy process for computers because of its visual complexity and differences between classes. However, Humans can simply identify a car by its logo or hood ornaments. Given the complexity of the problem, various approaches were used, starting from classical machine learning models to intricate deep learning models that achieved state-of-art results. Deep learning has been talked about a lot in recent years. And for a good reason, this subset of machine learning has imposed itself impressively in several research fields of which car classification was a part. Several methods and algorithms of deep learning were used to classify car make and model, primarily, Convolutional Neural networks (CNN), which are powerful programming models allowing in particular image recognition by automatically assigning each image provided as input a label corresponding to its class. Also, deep Neural Networks(DNN) a multilayer neural networks which can include millions of neurons, divided into several dozen layers. Deep learning empowers Artificial Intelligence to learn new rules to be more reliable and efficient. The exponential improvement in computing power and the development of related applications allow artificial intelligence to generate more complex and dense layers of neurons.

The challenge with model and car classification is the fine-grained feature. Compared to basic image recognition, the dataset is more diverse in contrast to the similarities found pixel-wise. The question is whether a model can differentiate between different cars model and make found on the fine-grained dataset. To solve this issue, a deep learning model should be able to adapt and recognize the similarities found in the dataset and enhance the prediction. Hence, a compound scaling model in which width, depth, and resolution are scaled so that the model captures more fine-grained patterns. In this paper, we present a novel approach to classifying models and car makes. Inspired by the recent work on scaling neural networks, we worked with the EfficientNet [9] model pre-trained on the ImageNet dataset. We fine-tuned the model to our needs, thus, adding layers to reduce complexity. We also used transfer learning for prior knowledge of the model weights. With such behavior, the model can adapt to different scenarios, therefore, better prediction accuracy. We conducted experiments on the challenging Stanford Cars dataset [10], which contains 196 different categories of cars taken from different angles. We have used FLASK API to create a web interface so as to achieve real-time prediction.

The paper is repartitioned as follows. First, we present a literature review on the subject of vehicle classification. Then we discuss the methodology in which we formulate the problem, explain the architecture used and discuss the image preprocessing, data augmentation, and transfer learning phase. Following is the experimentation section, where we discuss the dataset used, the model implementation, and the training loss function. After that, the results section is where we discuss the obtained results. Finally, a conclusion and perspective section.

II. RELATED WORK

Car make and model classification problem has widely been addressed using two research category methods. The first method focuses on handcrafted feature engineering. The second one, instead, focuses on machine learning and deep learning techniques. Since our method is based on a deep learning architecture, we mainly focus on related work with similar approaches. Xingyang Ni et al [11] 2021, compared two methods of car models and makes classification: a straight-forward classification and a more flexible metric learning method. They built their model based on the ResNet 50 pre-trained on the ImageNet dataset and retraining it on the VERI-WILD dataset that contains approximately 0.4 million images with 14 types of vehicles, and 149 different car makes. As for the result, upon using a cross-entropy loss, they found a 96.8% accuracy rate on the type classification and 95.6% on the make classification. While for the triplet loss method, they found an accuracy rate of 97.4% on the type classification and 95.3% on the make classification. In addition, they used a lifted structured loss method in which they found 97.7% on the type classification and 96.2% on the model classification. Ye Xiang et al [12] 2019, proposed a four-stage pipeline that consists of part detection, part assembling, topology constraint, and classification for fine-grained vehicle recognitions. They used a backbone model trunked at the middle in the first stage. Afterward, they called for pointwise convolutional layers that put together related parts into the same feature map. Eventually, the topology constraint covers depth wise convolutional layers and approximates the possibility of the topology correlation between associated parts. Finally, they evaluated the model on two public datasets: the Stanford Cars dataset and the CompCars dataset. In both datasets, various car viewpoints can be found. The results obtained were 94.3% accuracy on the CompCars dataset, 94.3% on the Stanford Cars dataset for the model classification and 99.6% for the make classification. Rachmadi et al. [13] 2018, proposed a pseudo-long short-term memory classifier for identifying a single image. The presented technique considers the split pictures to be time-series frameworks. Those images are outlined by cropping input images with a two-level spatial pyramid region configuration given to the P-LSTM classifier in a cycle. And to calculate the prediction of each class, they added a fully connected layer. They used the MIO-TCD dataset, which contains 648,959 vehicle images of 11 types of vehicles. They obtained a 97.98% accuracy rate. Jung et al. [14] 2017, trained ResNet models using actual traffic surveillance recordings. A joint fine-tuning method is employed to fine-tune all parameters and not only the final dense layer. They used DropCNN that arbitrarily drops the probabilities from the aforementioned backbones during training. They used the MOI-TCD dataset. They obtained a 97.9% accuracy rate. Hu et al [15] 2017,

presented a spatially weighted convolutional neural network that accommodates a predefined amount of pooling channels. The model then takes out deep convolutional neural network features with the enlightenment of its learned masks. They have achieved an accuracy of 93.1% on the Stanford Cars dataset and 97.6% on the CompCars dataset. Lee and Chung [16] 2017, proposed twelve local expert networks and six global networks. They used three neural network structures: AlexNet, GoogLeNet, and ResNet18. The local expertise and global networks are trained with the particular subsets and entire training set, respectively. They generated the prediction by combining the predictions of one local expert network and multiple global networks. They used the MIO-TCD dataset to get an accuracy rate of 97.92%. Huttunen et al. [17] 2016, presented a deep learning neural network that employed SVM (Support vector machine) on a dataset that contains over 6500 images. They found an accuracy rate of 97.75% for the deep learning method and 96.19% for the SVM method. Dong et al. [18] 2015 proposed a sparse Laplacian filter learning method to minimize the parameters of convolutional layers with a large number of unlabeled samples. They collected the BIT-Vehicle dataset, which contains 9850 high-resolution vehicle frontal-view images. They achieved an accuracy rate of 96.1%. Yang et al. [19] 2015, proposed a model based on pre-trained weights of the ImageNet dataset and fine-tuned it with the CompCars dataset. The result obtained was 82.9% accuracy in the car make, 76.7% in the car model, and 80.8% in car parts. Girshick et al. [20] 2014 proposed the fastest model, taking approximately 2 seconds for an object to be detected. The approach used similar layers for both the detection and classification tasks. First, the model detects the spatial geometric position of an object using a sliding window method. This allowed the model to classify vehicles accordingly, utilizing the image's extracted objects/features). The final accuracy of the model was a fascinating 73.2% in image classification. Although the classification accuracy wasn't that high, the model was fast enough to compensate for the lack. Wang et al. [21] 2013 used SPM (Spatial Pyramid Matching). The method mainly focuses on detecting the spatial distance that can be found between detected objects of an image. In addition to SPM, SIFT was used to extract features, followed by an LLC (Locality-constrained Linear Coding) to extract locations. The classification task achieved a 59.3% accuracy, improved by an SVM classifier later on. Krizhevsky et al. [22] 2012 proposed a low-level features extraction method using Gabor filters. The model was composed of higher layers that deal with classification tasks and lower layers that deal with extracting features, the classifier used was an SVM. The model performed a 93.3% accuracy on image identification and 83.3% on image classification. Cheung et al. [23] 2008 Used SIFT algorithm (ScaleInvariant Feature Transformation), the model matches interest points in car images. The framework consists of an optimization network that uses the geometry of the image to spot interest points in cars. If the matched points in the training model are similar to the test phase, those two points are called inliers; thus, the classification of the car matches. The only drawback of this framework is that it matches images in the dataset for the same angle only, resulting in a mismatch if the angle is modified. Bay et al. [24] 2008 developed an unsupervised learning model that acts on the behavior of labeled image subcategories. These subcategories were generated using a segmentation. The

model allowed only focus on essential image features. Hence, removing the background as it's considered as noise, the extracted features are then used for the classification task using a categorical loss function. This method led to the foundation of segmentation in car and model classification. Many methods have developed a fine-tuned model using segmentation filters. Some models even allowed modification in terms of kernel density. This ability make the model more robust in detecting and filtering important car image features. In recent work on vehicle classification, compound scaling has been used to extract fine-grained features. In our work, we sought to explore more the use of transfer learning with EfficientNet compound scaling coefficients pre-trained on the ImageNet and the MobileNet model architecture to classify model and car make.

III. METHODOLOGY

We describe our method as displayed in Fig. 1. The framework consists of three phases, the preprocessing and data augmentation phase, the model implementation phase, and the transfer learning phase. The model is trained end-to-end, and we developed a web interface for real-time prediction using the trained model weights. We will discuss each of these phases in the following sub-sections.

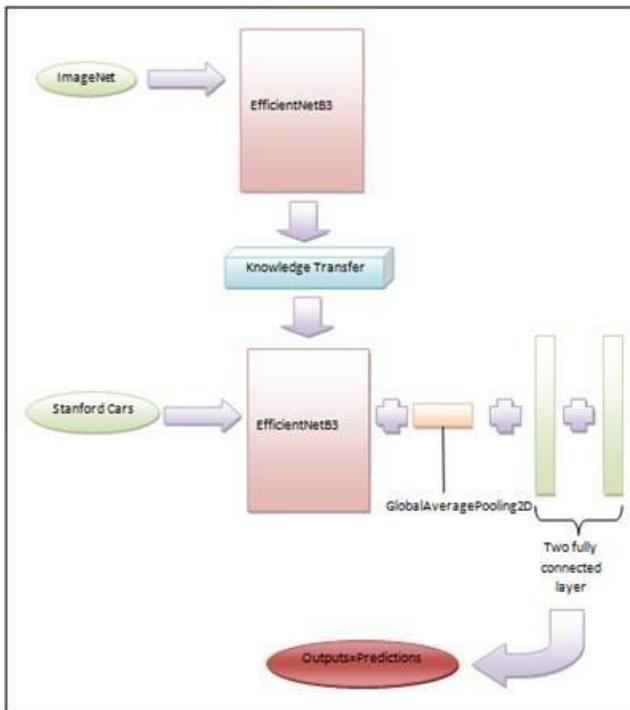


Fig. 1. The Training Workflow of our System Combining Transfer Learning and EfficientNetB3 Pre-Trained Model

A. Problem Formulation

We define the problem as a categorical classification scenario in which we ought to classify model and car make according to the scalability of the convolution neural network in question—That is, the learning behavior. Therefore we can

define a Convolution network as:

$$N = \bigodot_{i=1..s} \rho_i^{L_i}(X_{\langle H_i, W_i, C_i \rangle}) \quad (1)$$

where:

N defines a list of composed layers.

ρ denotes layer ρ_i is repeated L_i times in i .

$\langle H_i, W_i, C_i \rangle$ denotes the shape of tensor X where H_i , W_i are the spatial dimension and C_i is the channel dimension.

The objective is to find the best ρ , yet we want our model to be scalable in order to extract fine-grained features, so instead of focusing on finding the best ρ , we focus on finding the best scaling dimensions. As described in [9], the model fixes ρ and uniformly explores all layers parameters with a constant ratio. As such, we have an optimization problem, which can be formulated as follows:

$$\max_{d,w,r} N = \bigodot_{i=1..s} \hat{\rho}_i^{d \cdot L_i}(X_{\langle r \cdot \hat{H}_i, r \cdot \hat{W}_i, w \cdot \hat{C}_i \rangle}) \quad (2)$$

where:

d , w and r are the coefficients, depth, width and resolution respectively.

$\langle r \cdot \hat{H}_i, r \cdot \hat{W}_i, w \cdot \hat{C}_i \rangle$ defines the predefined parameters multiplied by the coefficients.

The advantage of scaling optimal d, w, r is that when scaling depth (d), the network tends to capture more complex features. Scaling the width (w) will allow the network to capture more fine-grained features. For the resolution (r), the network will have the ability to capture different patterns due to the enlargement of the resolution, making it easy to extract fine-grained features. With this in mind, the issue persists when maximizing the accuracy in contrast to the scalable parameters. The network will often become challenging as the scaling values depend on each other. Hence, we use EfficientNet with compound scaling parameters. Intuitively, the network scale is according to a compound coefficient determined by a grid search. This allows the network to fine-tune itself according to the optimal need. Fig. 2 shows the scaling behavior of EfficientNet compared with different other methods.

B. Model Architecture

We describe our model as a set of a combination between EfficientNet for compound scaling and MobileNet [25] as the model architecture. EfficientNet is a convolutional neural network that relies on scaling the width, depth, and resolution uniformly. In addition to that, the network has a small number of parameters compared to other models: it has only 12,320,535 parameters, but it has proven to reach better results on the ImageNet dataset compared to other models with a higher number of parameters. Thus, we transfer knowledge of the trained EfficientNet model and use it in our system using transfer learning. Fig. 3 shows the EfficientNet architecture. As for the MobileNet architecture, it uses mobile inverted bottleneck convolution (MBCConv), applying depth-separable convolution with residuals.

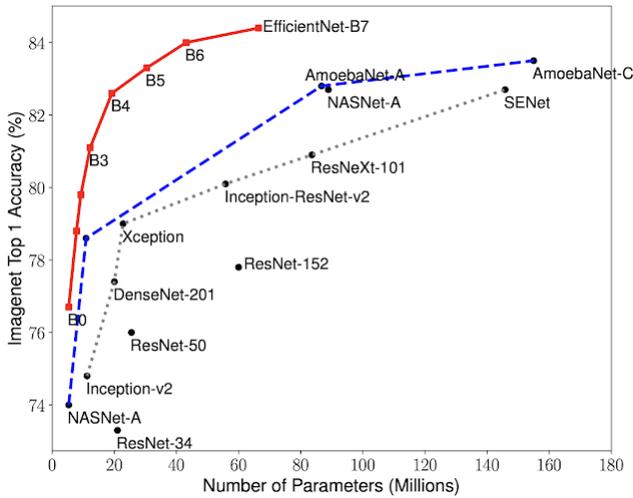


Fig. 2. Comparison between EfficientNets and other Existing CNNs on ImageNet

The main difference between the regular residual block and the inverted residual block is that the latter follows a narrow > wide > narrow approach. In contrast, the first follows a wide > narrow > wide path approach. For example, Fig. 4 shows the difference between Residual and Inverted residual blocks.

C. Image Preprocessing and Data Augmentation

Our dataset has a small number of cars in each class: about forty pictures. Training a deep neural network on few images is often challenging: the model having access to only a limited number of observations will tend to Overfit. In this case, the performance is poor on the test set while they are good on the training set. This phenomenon is often solved by increasing the size of the dataset and/or reducing the number of model parameters. The first method is often challenging to set up because the work of collecting/labeling new observations is laborious. The second possibility is conceivable for an image recognition problem. However, even the most miniature complex models can contain hundreds of thousands of parameters, which are tricky to achieve. As data augmentation allows new labeled images to be generated from those already available, it is a relatively more straightforward solution to implement, and the results can be surprising. The most well-known technique of data augmentation is image data augmentation. It combines the methods used to artificially increase the size of a training dataset by creating modified versions of images from the available training images. We can then effectively improve the learning process as it results in more training samples for the neural network model. Augmentation techniques can create image variations that can enhance the ability of training models to generalize what they have learned to new images, which significantly improves model performance. The data augmentation applies only to the training set and not the validation or test set. This differs from data preparation, such as image resizing, which must be performed consistently across the entire dataset interacting with the model. Fig. 5 shows a sample of data augmentation on the Stanford Cars dataset.

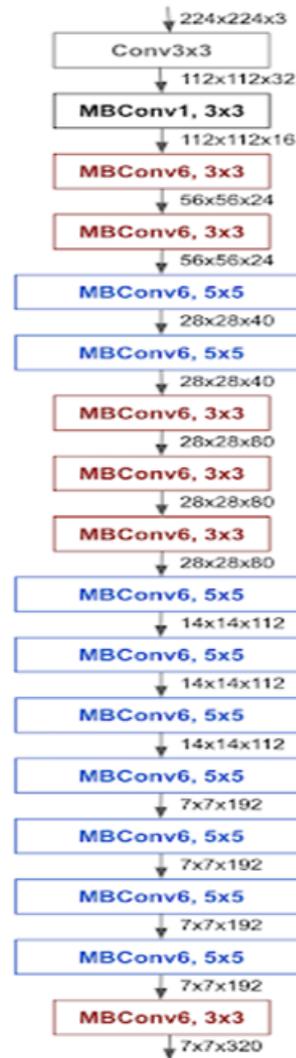


Fig. 3. EfficientNet Architecture

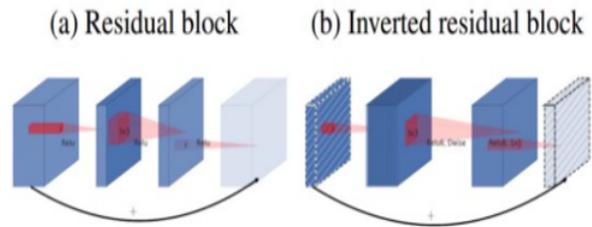


Fig. 4. Comparison between Residual Block and Inverted Residual Block

D. Transfer Learning

Transfer learning [26] has become common in the past few years because it has proven to achieve better results even with the use of a small amount of data. In our work, we have used transfer learning, a supervised learning technique that consists of taking a pre-trained model and reusing it on another dataset. Fig. 6 shows the workflow of the transfer learning technique.

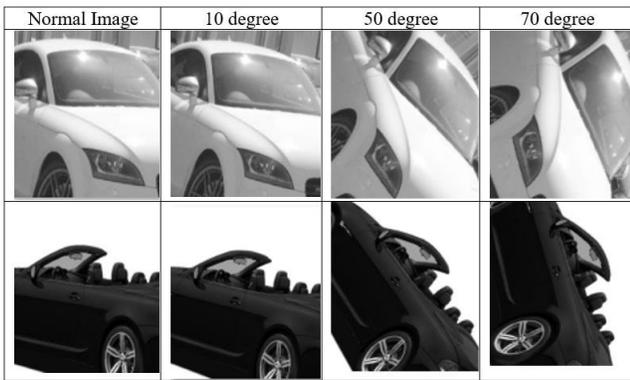


Fig. 5. Augmentation Technique Implemented on the Stanford Cars Dataset

Transfer Learning consists of the transfer of knowledge from one task to another. This behavior allows the network to solve similar problems with the same pre-trained model. This will eventually improve the quality of learning and reduce the computation time. However, deep learning requires always having a large dataset to operate the neural networks at their foremost. Therefore, we can adjust using Transfer Learning to get better predictions even with a small dataset.

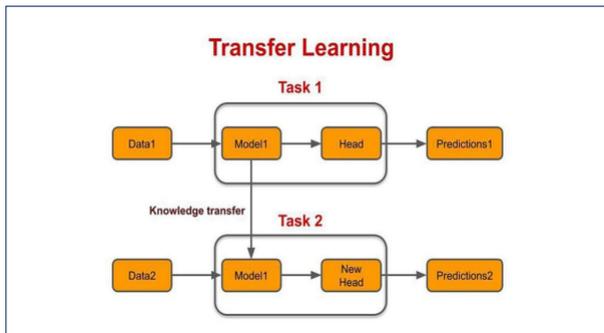


Fig. 6. Transfer Learning Workflow

IV. EXPERIMENTATION

A. Understanding the Stanford Dataset

Stanford Cars dataset has been widely used for model and car classification. The dataset was collected in 2013 and contained 196 different categories of cars (model, make, and year). The dataset is split into 8,144 pictures for the train set and 8,041 pictures for the test set. Fig. 7 shows an example of some quantitative car images in the Stanford dataset.

The files anno-test.csv and anno-train.csv are a table of six columns. The first column represents the name of the picture. From the second to the fifth column, we have four values of pixels that show the exact emplacement of the vehicle in the image. The last column contains information that classifies the car. The list of all classes is in the "classes.csv" file. Fig. 8 shows the dataset files structure.



Fig. 7. An Overview of the Diversity in the Stanford Dataset

File Name	Column 1	Column 2	Column 3	Column 4	Column 5	
anno_test.csv	00001.jpg	30	52	246	147	181
	00002.jpg	100	19	576	203	103
	00003.jpg	51	105	968	659	145
	00004.jpg	67	84	581	407	187
anno_train.csv	00001.jpg	39	116	569	375	14
	00002.jpg	36	116	868	587	3
	00003.jpg	85	109	601	381	91
	00004.jpg	621	393	1484	1096	134

Fig. 8. Data and csv Files Structure

B. Model Implementation

Car classification is a challenging task in machine learning due to the variety of details in each car. Therefore, we used transfer learning on the EfficientNet model pre-trained on the ImageNet dataset and then fine-tuned the model to get better results. First, we started by adding layers to our base model, mainly the globalAveragePooling2D layer, to reduce the variance and complexity of calculations, two fully-connected layers with the activation function "relu," and the integration of dropout to reduce overfitting. Next, we trained our model on the Stanford cars database combined with MoVITS Dataset [27]. Finally, to get a higher accuracy rate, we fine-tuned the model by unfreezing our entire model and retraining it.

C. Training on the Stanford Dataset

We used adam as an optimizer and categorical cross-entropy as a loss function to train our model. The use of adam, in this case, is due to the quick convergence that the optimizer allows. Since the dataset is quite complex in term of diversity, adam will help reduce computation time and converges in a significantly lower period. Due to the different classes in the dataset, we chose categorical cross-entropy as our loss function. To evaluate our model, we minimize the loss and compute the accuracy of the training and validation data. Since this is a classification task, we also demonstrate a confusion matrix to help visualize the behavior of the network. We define the categorical cross-entropy loss function as follows:

$$\mathcal{L}(y, y') = \sum_{j=0}^M \sum_{i=0}^N (y_{i_j} * \phi) \quad (3)$$

where:

$$\phi = \log (y'_{ij}).$$

y' is the predicted value, y is the ground-truth value.

To compute the accuracy, we use the following principle:

$$A = \frac{TP + TN}{TP + TN + FP + FN} \quad (4)$$

where:

A represents the Accuracy.

TP + TN represents the number of correct predictions.

TP + TN + FP + FN represents the total number of predictions.

In summary, anytime the prediction is incorrect, the forecast is False. Otherwise, it is True. Therefore, the final objective is to maximize the prediction as True (True Positive and True Negative) and minimize the prediction as False (False Positive and False Negative).

V. RESULTS AND DISCUSSION

Compared with ResNet implementation in [11], we show that using EfficientNet has significantly enhanced the network's ability to extract more detailed features of the dataset. Adding data augmentation, the model learns to adapt to different image perspectives making it more robust on unseen data. Furthermore, in contrast to [12], using pointwise convolutional layers to extract fine-grained features, we incorporated scaling coefficients. This allowed the network to scale parameters for an optimal state. Adding MobileNet as a model architecture with mobile inverted bottleneck convolution reduced memory requirement compared to classical residual block.

We trained our model first using 44 epochs. Comparing the training and test loss, as shown in Fig. 9, the model started to find the optimal state using a grid search for the model's scaling coefficient at around 20 epochs. After 20 epochs, the model fluctuated, considering the complexity and deviation of the dataset combined with the augmented images. Finally, After 40 epochs, the model stagnates. Thus, we deduct that the model achieved an optimal state for the given parameters. We found an accuracy of 88% on the train set and 82% on the test set. Fig. 9 shows the results.

After the first experiment, we understood the behavior of our model, especially after 20 epochs where more fine-grained features are extracted due to the compound scaling of the optimal d, w , and r in the network. To enhance our model's accuracy, we fine-tuned it by retraining it to only 24 epochs where the network understands mostly essential image features. Compared to the previous experiment. We achieved an accuracy of 95% on the train set and 90% on the test. Fig. 10 shows the results.

To showcase the confusion of the network with respect to the predicted values, we generate a confusion matrix which is a summary of the results of predictions about a classification problem to visualize our prediction better. Fig. 11 shows the confusion matrix obtained. Correct and incorrect predictions are highlighted and broken down by class. The results are thus compared with the actual values. This matrix helps

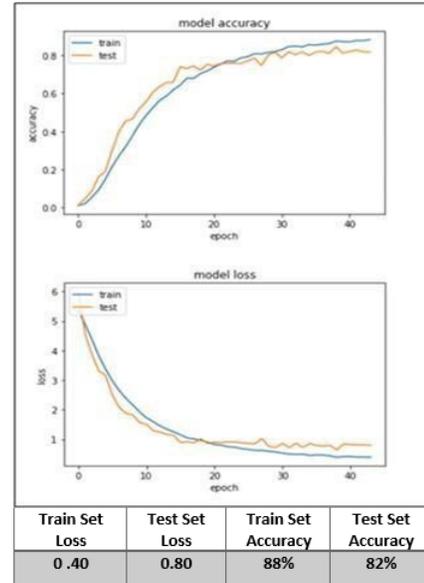


Fig. 9. Model Accuracy and Loss before Fine-Tuning

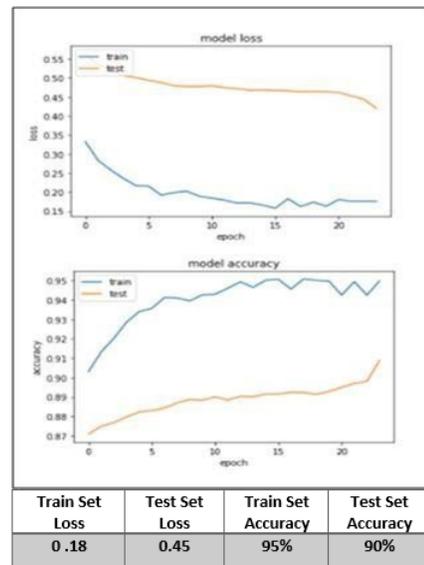


Fig. 10. Model Accuracy and Loss after Fine-Tuning

understand how the classification model is confused when making predictions.

The model accuracy on unseen data is significantly interesting, especially considering the use of augmented data. This allowed the model to understand different image perspective views and find similarities in fine-grained features to be then able to label the correct and incorrect predictions correctly. At this stage, our model's weight is ready to be used for real-time prediction.

Now that we have our model trained, we developed a friendly interface for real-time prediction using the FLASK API and the weighted model trained. The workflow of the API is described in Fig. 12.

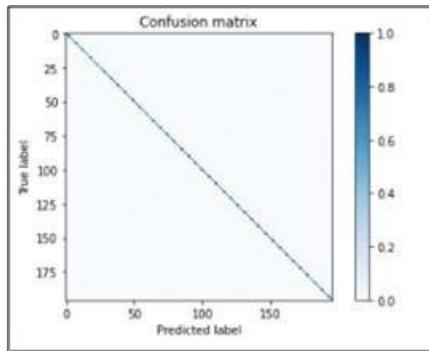


Fig. 11. Confusion Matrix of the Classification Task

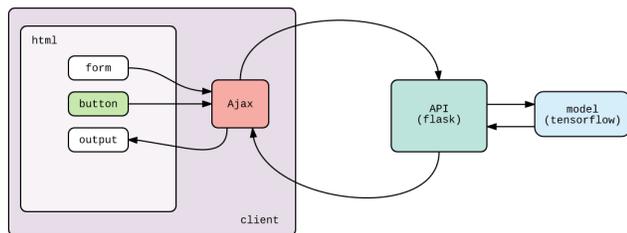


Fig. 12. FLASK API General Workflow

Our final interface is composed of an upload section where the user can freely upload a car image. Upon clicking on submit, the network uses the trained model to predict based on the model weights and outputs the predicted label of the given car image, describing the model and car make. Fig. 13 shows an example of the prediction mechanisms.

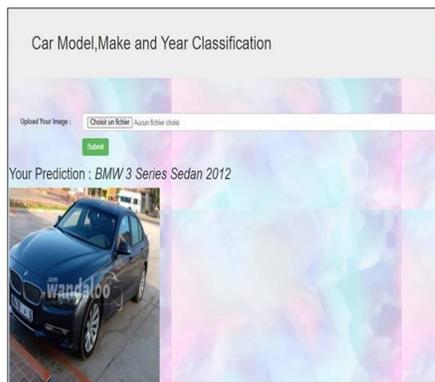


Fig. 13. An Example of Real Time Prediction in our Web Interface using Flask API

VI. CONCLUSION AND PERSPECTIVES

In this paper, we explored the use of transfer learning with EfficientNet compound scaling coefficients pre-trained on the ImageNet and the MobileNet model architecture to classify model and car make. The use of such a combination proved to be efficient in this task. We used the Stanford Cars dataset and fine-tuned the model to find an accuracy rate of 90%. Additionally, we have implemented a web interface to predict car images in real-time using Flask API. Extracting

fine-grained features is indeed a complex task, yet, we show that combining and fine-tuning the model can significantly enhance accuracy. Furthermore, we can improve the project prediction by building a system that identifies a car's plate number using Optical Character Recognition (OCR), which converts digital images to electronic text. The OCR output is an ASCII code that contains the text of the license plate and which can be compared to existing databases containing additional information on the car owner, such as his issuance badge, serial number, etc. These extracted features will then be used to improve the accuracy of the overall model.

ACKNOWLEDGMENT

This work was supported in part by the CNRST and in part by the MESRSFC, through the Development of an Integrated System for Traffic Management and Detection of Road Traffic Infractions Project.

REFERENCES

- [1] A. Saif and Z. R. Mahayuddin, "Robust drowsiness detection for vehicle driver using deep convolutional neural network," *International Journal of Advanced Computer Science and Applications*, vol. 11, no. 10, 2020.
- [2] H. Ait Abdelali, O. Bourja, R. Haouari, H. Derrouz, Y. Zennayi, F. Bourzeix, and R. Oulad Haj Thami, "Visual vehicle tracking via deep learning and particle filter," in *Advances on Smart and Soft Computing*. Springer, 2021, pp. 517–526.
- [3] H. A. Abdelali, H. Derrouz, Y. Zennayi, R. O. H. Thami, and F. Bourzeix, "Multiple hypothesis detection and tracking using deep learning for video traffic surveillance," *IEEE Access*, vol. 9, pp. 164 282–164 291, 2021.
- [4] H. Derrouz, A. Elbouziady, H. Ait Abdelali, R. Oulad Haj Thami, S. El Fkihi, and F. Bourzeix, "Moroccan video intelligent transport system: Vehicle type classification based on three-dimensional and two-dimensional features," *IEEE Access*, vol. 7, pp. 72 528–72 537, 2019.
- [5] H. Derrouz, A. Cabri, H. Ait Abdelali, R. Oulad Haj Thami, F. Bourzeix, S. Rovetta, and F. Masulli, "End-to-end quantum-inspired method for vehicle classification based on video stream," *Neural Computing and Applications*, pp. 1–16, 2022.
- [6] O. Bourja, K. Kabbaj, H. Derrouz, A. El Bouziady, R. O. H. Thami, Y. Zennayi, and F. Bourzeix, "Movits: Moroccan video intelligent transport system," in *2018 IEEE 5th International Congress on Information Science and Technology (CiSt)*, 2018, pp. 502–507.
- [7] O. BOURJA, H. DERROUZ, H. A. ABDELALI, A. MAACH, R. O. H. THAMI, and F. BOURZEIX, "Real time vehicle detection, tracking, and inter-vehicle distance estimation based on stereo vision and deep learning using yolov3," *International Journal of Advanced Computer Science and Applications*, vol. 12, no. 8, 2021. [Online]. Available: <http://dx.doi.org/10.14569/IJACSA.2021.01208101>
- [8] O. Bourja, A. Maach, Y. Zennayi, F. Bourzeix, and T. Guerin, "Speed estimation using simple line," *Procedia Computer Science*, vol. 127, pp. 209–217, 2018.
- [9] Q. V. L. Mingxing Tan, "Efficientnet: Rethinking model scaling for convolutional neural networks," *ICML*, 2019.
- [10] J. Krause, M. Stark, J. Deng, and L. Fei-Fei, "3d object representations for fine-grained categorization," in *Proceedings of the IEEE international conference on computer vision workshops*, 2013, pp. 554–561.
- [11] X. Ni and H. Huttunen, "Vehicle attribute recognition by appearance: Computer vision methods for vehicle type, make and model classification," *Journal of Signal Processing Systems*, vol. 93, no. 4, pp. 357–368, 2021.
- [12] Y. Xiang, Y. Fu, and H. Huang, "Global topology constraint network for fine-grained vehicle recognition," *IEEE Transactions on Intelligent Transportation Systems*, vol. 21, no. 7, pp. 2918–2929, 2019.
- [13] R. F. Rachmadi, K. Uchimura, G. Koutaki, and K. Ogata, "Single image vehicle classification using pseudo long short-term memory classifier," *Journal of Visual Communication and Image Representation*, vol. 56, pp. 265–274, 2018.

- [14] H. Jung, M.-K. Choi, J. Jung, J.-H. Lee, S. Kwon, and W. Young Jung, "Resnet-based vehicle classification and localization in traffic surveillance systems," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2017, pp. 61–67.
- [15] Q. Hu, H. Wang, T. Li, and C. Shen, "Deep cnns with spatially weighted pooling for fine-grained car recognition," *IEEE Transactions on Intelligent Transportation Systems*, vol. 18, no. 11, pp. 3147–3156, 2017.
- [16] J. Taek Lee and Y. Chung, "Deep learning-based vehicle classification using an ensemble of local expert and global networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2017, pp. 47–52.
- [17] H. Huttunen, F. S. Yancheshmeh, and K. Chen, "Car type recognition with deep neural networks," in *2016 IEEE intelligent vehicles symposium (IV)*. IEEE, 2016, pp. 1115–1120.
- [18] Z. Dong, Y. Wu, M. Pei, and Y. Jia, "Vehicle type classification using a semisupervised convolutional neural network," *IEEE transactions on intelligent transportation systems*, vol. 16, no. 4, pp. 2247–2256, 2015.
- [19] L. Yang, P. Luo, C. Change Loy, and X. Tang, "A large-scale car dataset for fine-grained categorization and verification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3973–3981.
- [20] T. D. Ross Girshick, Jeff Donahue and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," *CVPR*, 2014.
- [21] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong, "Locality-constrained linear coding for image classification," *Onternational Conference on Human Computer Interactions (ICHCI)*, 2013.
- [22] G. E. H. Alex Krizhevsky, Ilya Sutskever. "Imagenet classification with deep convolutional neural networks," *Advances in Neural Information Processing Systems 25 (NIPS)*, 2012.
- [23] S. Cheung and A. Chu, "Make and model recognition of cars," *Projects in Vision and Learning*, 2008.
- [24] H. Bay, T. Tuytelaars, and L. Gool, "Surf: Speeded up robust features," *Computer Vision and Image Understanding*, pp. 346–359, 2008.
- [25] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," *Computer Vision and Pattern Recognition (cs.CV)*, 2018.
- [26] N. Donges, "What is transfer learning? exploring the popular deep learning approach," *Built In*, 2019.
- [27] H. Derrouz et al., "The moroccan video intelligent transport system dataset for vehicle detection," 2021. [Online]. Available: <https://data.mendeley.com/datasets/5jcg5vfx58/3>

Cache Complexity of Cache-Oblivious Approaches: A Review and Extension

Inas Abuqaddom, Sami Serhan, Basel A. Mahafzah
Department of Computer Science
King Abdullah II School of
Information Technology
The University of Jordan
Amman, Jordan

Abstract—The latest direction in cache-aware/cache-efficient algorithms is to use cache-oblivious algorithms based on the cache-oblivious model, which is an improvement of the external-memory model. The cache-oblivious model utilizes memory hierarchies without knowing memories' parameters in advance since algorithms of this model are automatically tuned according to the actual memory parameters. As a result, cache-oblivious algorithms are particularly applied to multi-level caches with changing parameters and to environments in which the amount of available memory for an algorithm can fluctuate. This paper shows the state of the art in cache-oblivious algorithms and data structures; each with its complexity concerning cache misses, which is called cache complexity. Additionally, this paper introduces an extension to minimize the cache complexity of neural networks by applying an appropriate cache-oblivious approach to neural networks.

Keywords—Cache complexity; cache-oblivious algorithm; memory hierarchy; neural network

I. INTRODUCTION

The processor speed is much faster than the main memory speed. The impact of this hole can be decreased by utilizing a hierarchy of multi-level caches in an effective way between the processor and the main memory [1]. Thus, modern computers have a memory hierarchy to speed up accessing memory, such that the speed of accessing a memory level becomes slower as its size becomes larger. Table 1 shows an example of a memory hierarchy. In most processors, the level 1 cache (L1) is on the same chip as the CPU, whereas the level 2 cache (L2) is on a separate chip [2], [3], [4].

TABLE I. MEMORY HIERARCHY

Memory level	Size	Response time
CPU registers	around 100B	around 0.5ns
L1 Cache	around 64KB	around 1ns
L2 Cache	around 1MB	around 10ns
Main memory	around 2GB	around 150ns
Hard disk	around 1TB	around 10ms

The response time increases as the memory size increases, as shown in Table I, consequently, there is an inverse relationship between memory size and its speed; as the memory level becomes larger, its speed becomes slower. Generally, each memory level communicates directly with the slower directly connected memory level. Additionally, data is transferred in blocks to reduce the effects of slow access [5]. For example,

to read one element of an array, the main memory will additionally transmit a "block" of consecutive words. Thus, accessing the other words of the transferred block is free in terms of memory transfers. The design of a multi-level memory hierarchy is more complex compared with single-level memory; such as extra design decisions are needed for each level of the memory hierarchy [6].

One performance metric of an algorithm is called memory performance, which measures the utilization of the memory hierarchy in an algorithm. Thus, algorithms have to know the memory hierarchy, memory size, and block size to achieve high memory performance; these algorithms are called cache-aware algorithms or cache-efficient algorithms. Nevertheless, cache-aware algorithms reduce running time up to 50% compared with other algorithms. On the other hand, algorithms, that worry about memory performance, have to be tuned according to the underlying cache size. The big question here is: "Can we design algorithms to efficiently utilize memory hierarchy, without knowing the underlying cache size?" The answer is yes by using cache-oblivious algorithms [7].

Neural networks greedily consume the memory hierarchy, hence there is a strong need to design neural network algorithms efficiently in terms of memory performance [8], [9]. In this paper, we show the recent direction of the memory model, which is the cache-oblivious model, and the state of the art in cache-oblivious algorithms and data structures. Furthermore, we applied the appropriate cache-oblivious approach to neural networks to improve their memory performances.

The rest of this paper is organized as follows: Sections II and III explain the cache-oblivious model and cache complexity, respectively. Section IV discusses a set of cache-oblivious algorithms, while Section V introduces the extension, which is cache-oblivious neural networks. Finally, we provide a summary of the best lately known cache-oblivious algorithms in Section VI.

II. CACHE-OBLIVIOUS MODEL

The literature discusses two memory-hierarchy models; the external-memory model and the cache-oblivious model. The memory hierarchy of the external-memory model consists of two levels; cache and disk. Such that, cache refers to the near memory to the CPU as disk refers to the far memory from the CPU. The basic unit of transferring data between two memory levels is a block, so the memory is divided into blocks of B

words (a certain number of bytes). At most, the memory of size M can store $\frac{M}{B}$ blocks from data of size N words [10], [11].

To reduce the number of transferring blocks, cache-aware algorithms are designed based on the external-memory model; such algorithms worry about the existence of two levels of memory and their parameters. Cache-aware algorithms perform efficiently with two levels of memory, in terms of memory transfers. But in reality, memory hierarchy is more than just two levels of memory. To cope with this fact, the cache-oblivious model is introduced as an extension to the external-memory model [12], [11].

The cache-oblivious model utilizes all cache levels efficiently without tuning, which means algorithms can utilize the memory without knowing its size M and block size B . Cache-oblivious model is the best choice for the environment that provides a fluctuation amount of available memory for an algorithm. Better algorithms for this model can be better for any possible values of M and B . Cache-oblivious algorithms are particularly helpful for multi-level caches and for caches with changing values of M and B . Recently, the direction of designing cache-efficient algorithms is cache-oblivious algorithms. The advantages of cache-oblivious algorithms are as follows [13], [14]:

- 1) Inclusion: As the cache-oblivious algorithms optimally progress between two adjacent levels of the memory hierarchy, which are cache referring to the near memory to the CPU and disk referring to the far memory from the CPU. Then, these algorithms are automatically adapted between any two adjacent memory levels with different values of M and B , because they are not fixed in cache-oblivious algorithms.
- 2) Constant optimal factor: Optimality means the minimum number of cache misses for an algorithm. There is no way to reduce the cache misses for an algorithm less than its optimal cache misses. When the number of memory transmissions or cache misses is optimal to a constant c between any two adjacent memory levels, then this optimality is kept within a weighted factor between the other two adjacent memory levels, such that the weighted mixture will be corresponding to the relative speeds of the memory levels. For example, assuming the optimal number of cache misses between two adjacent memory levels to traverse data of size N is $\left\lceil \frac{N}{B} \right\rceil$. Then, this factor is kept between the other two adjacent memory levels and is weighted by their relative speeds. Thus, algorithms in a two-level memory model can be designed and analyzed to gain outcomes for some levels of the memory hierarchy.
- 3) Self-tuning: Typical cache-aware algorithms need tuning to various cache parameters which are no longer on hand from the manufacturer and are often hard to extract automatically. Parameter tuning makes code portability difficult, while cache-oblivious algorithms perform well on all machines without modifications based on the cache parameters.

III. CACHE COMPLEXITY

Cache complexity is the number of cache misses that are incurred by an algorithm for a problem of input size N and denoted by $T(N)$. The transfer unit between two adjacent memory levels is a block of size B words to amortize the access time cost. Typically, the main goal of an algorithm is to minimize the cache complexity $T(N)$, which is bounded by N as the upper bound and $\frac{N}{B}$ as the lower bound. In other words, the number of memory transfers at most equals the input size when each operation incurs a cache miss, whereas storing related elements in the same memory block B , which is called locality, reduces the number of cache misses into $\frac{N}{B}$ as a lower bound. We are concerned about complexity for large problem N when it is greater than B or even greater than M [15].

Typical cache complexity is a function of N and B because the minimum unit of transferred data is B with one cost unit. Also, M is relevant in cache complexity especially for algorithms with recursion when data fits in the cache and has been loaded in it, then the accessing cost will be zero. Generally, to compute the cache complexity of divide and conquer algorithms, we have to enlarge the base case to fit either B or M sizes [16], [17].

IV. CACHE-OBLIVIOUS ALGORITHMS

Cache-oblivious algorithms are mainly concerned with the efficiency of fetching large data into memory, which needs many memory transfers. This Section shows various cache-oblivious algorithms and their lower bound of memory transfers. Generally, cache-oblivious algorithms provide the lower bound of memory transfers utilizing the ideal cache, which is based on the following assumptions [18], [19], [20]:

- Optimal page replacement, such as evicting the least-recently-used block (LRU) or evicting the oldest-used block (FIFO) [21].
- Full associativity, such that the transferred memory block B can be stored at any available block in the cache.
- Tall cache, which means the number of blocks $\frac{M}{B}$ is greater than the block size B .

A. Scanning Approach

Scanning algorithms access all data items to perform some tasks such as finding maximum element, getting the average of elements, classifying elements, etc. Therefore, scanning algorithms touch all data items once and in the same order as they are stored, consequently, scanning algorithms are not aware of the cache size M . The cache complexity of a scanning algorithm is shown in Equation 1, such that N items lay out in contiguous blocks of memory. The ceiling function indicates one more memory transfer than $\frac{N}{B}$, because either N is not divided by B , or N does not start to lay out items from the beginning of a block. In other words, data items of size B require one memory transfer, but sometimes they require two memory transfers according to the alignment of the data items. For example, if the size of N is 8 and the size of B is 4, then scanning N incurs 2 memory transfers. However, if we did

not start to lay out N items from the beginning of a block as shown in Fig. 1, where N items are displayed in gray color. Even though the N size is $2B$, scanning N incurs 3 memory transfers instead of 2 because of the alignment [10].

$$T(N) = O\left(\left\lceil \frac{N}{B} \right\rceil\right) \quad (1)$$

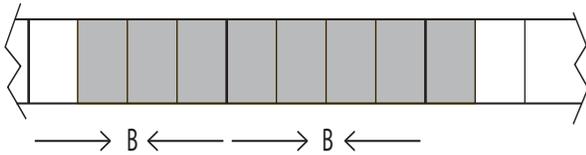


Fig. 1. Two Consecutive Blocks in Memory (the Gray Blocks Represent a Two-Block-Data item).

B. Divide and Conquer Approach

Divide and conquer algorithms recursively split problems into non-overlapping smaller sub-problems, solve them, and combine them. Divide and conquer algorithms recur down to the base case of constant size. In terms of memory transfers, we consider the base case that either fits one block B or fits within cache size ($\leq M$). Such a size is considered the critical place for memory transfers, where all the cost is. Because once data fits in the cache, the accessing cost will be zero. However, the base case below B is kind of trivial [6].

Nevertheless, the divide and conquer approach is a basic technique for designing cache-oblivious algorithms, which often afford optimal cache complexity within a constant factor among the levels of a memory hierarchy [6].

C. Search Tree Approach

Initially, we will discuss a binary search tree using a divide and conquer approach, where the base case is a sub-tree of size B [6]. Nevertheless, the cache-complexity of the binary search tree on a sorted array is:

$$T(N) = O(\lg N - \lg B) = O\left(\lg \frac{N}{B}\right)$$

Where $\lg N$ is the height of the binary tree, and $\lg B$ is the leaf height i.e. the base case height. However, B-tree can achieve the optimal cache-complexity i.e. $O\left(\frac{\lg N}{\lg B}\right)$ by making the branching factor value between B and $\frac{B}{2}$. The drawback of B-trees is that the branching factor cannot be tuned easily between any two memory levels. Nevertheless, if the branching factor is known for all levels of the memory hierarchy, then the optimal cache-complexity can be achieved by using a B-tree algorithm [6].

The authors of [22] introduced an efficient search tree algorithm that can be tuned easily between any two memory levels and their algorithm achieves the optimal cache complexity. This algorithm is the so-called cache-oblivious tree (or Van Emde Boas layout), which is a binary search tree, but each recursive sub tree is laid out in a single segment of memory. Accordingly, the tree is recursively split from the middle, so the height of the tree is $\lg N$, see Fig. 2. We keep splitting

each half recursively into two almost equal halves. At some point in this recursion, we reach halves of size less than or equal to B [22], [6].

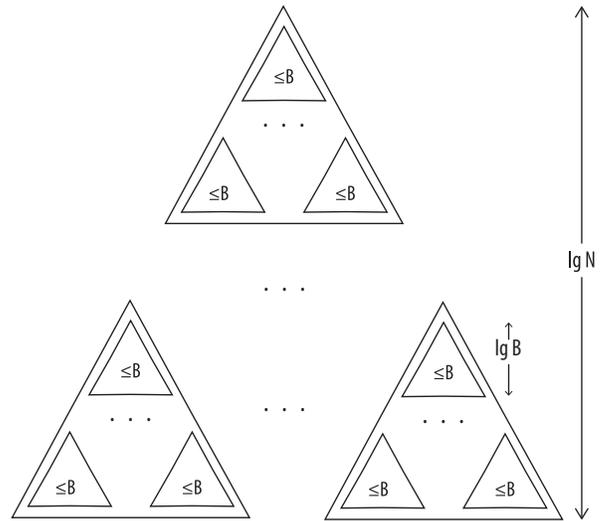


Fig. 2. Layout of the Cache-Oblivious Tree.

For analysis, following a root-to-leaf path visits some sequence of triangles, where each of them fits in, basically one block. In other words, triangle size is, at most B and at least \sqrt{B} , which is generated from splitting $B+1$ sub-tree into two further sub-trees, each of size \sqrt{B} . Thus, the height of the smallest triangle i.e. the leaf belongs to $[\lg \sqrt{B}, \lg B]$. Then, at most the number of visited triangles for a root-to-leaf path of at most $\lg N$ height is as follows [22]:

$$\frac{\lg N}{\lg \sqrt{B}} = \frac{\lg N}{\frac{1}{2} \lg B} = 2 \log_B N$$

Each block requires one transfer, but sometimes a block requires two transfers because of the alignment of each block, as shown in Figure 1. Therefore, the cache-oblivious tree incurs at most $2 \times (2 \log_B N)$ memory transfers. As a result, the cache-oblivious tree performs INSERT, DELETE, and SEARCH operations through an optimal cache complexity, which is shown in Equation 2 [22].

$$T(N) = O(4 \log_B N) \quad (2)$$

D. Sorting Algorithms

Sorting algorithms take N elements of some arbitrary order and put them into a sorted order. Nevertheless, sorting is an essential algorithm in computer science, since it can decrease the complexity of a problem, especially in searching and database problems. There are many ways to sort elements using various algorithms, such as bubble, selection, insertion, merge, quick, heap, radix, bitonic, and bucket sort [23], [24], [25], [26].

The obvious and easiest way to sort elements is by doing N inserts into a regular B-tree, which incurs $O(N \log_B N)$. B-trees are efficient for searches but are not efficient for very frequent updates [27]. To get a better result, an efficient sorting algorithm can be used such as merge sort, which uses a divide and conquer approach.

Merge sort divides the problem into two parts, recursively sorts each part, and recursively merges the two parts. Thus, merge sort requires three parallel scans; one for the first portion, another for the second portion, and the third scan for the merged array, i.e. the sorted array. Such that, we compare the first elements of both unsorted portions, output one of them into the merged array, and move the unsorted portion pointer to the next element, then compare, output one of them, and so on till reaching the end of both unsorted portions [28]. Accordingly, the whole block is scanned, kicked out, and then the next one is read, so the three parallel scans can be afforded, as long as, $\frac{M}{B} \geq 3$.

Nevertheless, we always have to be careful with the base case, whose best size is $\leq M$, because when a sub-array of size M is reached, the whole thing is read without incurring any more cost as long as, the sub-array stays within a region of size M . Figure 3 shows the recursion tree of merge sort, which is better than B-trees and incurs $O\left(\frac{N}{B} \lg \frac{N}{M}\right)$. In terms of memory transfers, $\frac{N}{B}$ is the number of sorted elements and $\lg \frac{N}{M}$ is the longest path through the recursion tree to sort an element [29].

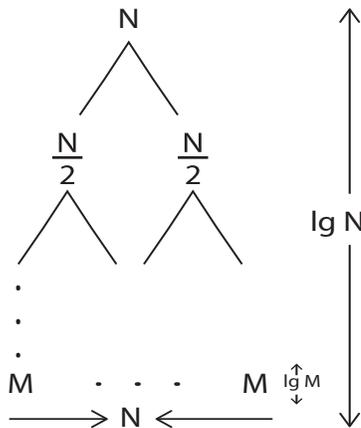


Fig. 3. The Recursion Tree of Binary Merge Sort.

Furthermore, to do a multi-way merge sort such as $\frac{M}{B}$ -way merge sort, then we can mimic the binary merge sort but with $\frac{M}{B}$ portions instead of two portions [30]. In other words, $\frac{M}{B}$ parallel scans are needed, so we have to lay out in cache the first blocks of these $\frac{M}{B}$ sub-arrays, whose size for each of them is $(N/\frac{M}{B})$. As a result, the same solution as binary merge sort is achieved but with a shorter longest path through the recursion tree as $(\log_{\frac{M}{B}} \frac{N}{M} + 1)$, such that:

$$\begin{aligned} \log_{\frac{M}{B}} \frac{N}{M} + 1 &= \log_{\frac{M}{B}} \left(\frac{N}{B} \frac{B}{M} \right) + 1 \\ &= \log_{\frac{M}{B}} \frac{N}{B} + \log_{\frac{M}{B}} \frac{B}{M} + 1 \\ &= \log_{\frac{M}{B}} \frac{N}{B} - \log_{\frac{M}{B}} \frac{M}{B} + 1 \\ &= \log_{\frac{M}{B}} \frac{N}{B} - 1 + 1 \\ &= \log_{\frac{M}{B}} \frac{N}{B} \end{aligned}$$

Accordingly, Equation 3 shows the cache complexity of sorting N elements using an $\frac{M}{B}$ -way merge sort, which is optimal. Also, an $\frac{M}{B}$ -way merge sort is the so-called cache-

oblivious sorting algorithm [29].

$$T(N) = O\left(\frac{N}{B} \log_{\frac{M}{B}} \frac{N}{B}\right) \quad (3)$$

E. Priority Queue Data Structure

A priority queue is a queue, where every item is associated with a priority. An item with the highest priority is dequeued before any other item. The main operation associated with the priority queue is getting the highest priority item at any given time. Priority queue provides INSERT and DELETE-MIN operations, to add an item to a queue and to dequeue the highest priority item from it, respectively, as the highest priority item corresponding to the minimum number. Literature introduces priority queues supporting a different set of operations. In this section, we discuss the cache-oblivious priority queue supporting INSERT and DELETE-MIN operations [31], [32], [33].

The cache-oblivious priority queue uses a bunch of arrays in a linear order instead of a bunch of B-trees. Because priority queue using arrays is simpler and incurs fewer memory transfers than priority queue using B-trees, which are the best choice for searching operations. The cache-oblivious priority queue using the divide and conquer approach is arranged in levels, as each level is decomposed of two types of buffers; one “up buffer” and a set of “down buffers” such that each buffer is of a certain size. In other words, the cache-oblivious priority queue levels are recursive smaller priority queues, where the highest priority item exists in the smaller priority queue, i.e. the smaller level, in the cache [34].

The cache-oblivious priority queue for N items has $\lg \lg N$ levels, whose sizes are arranged from top to bottom as $N, N^{2/3}, N^{4/9}, \dots, X^{9/4}, X^{3/2}, X, X^{2/3}, \dots, C$, respectively, as shown in Fig. 4. C is a constant size level, and X is a size between N and C . At the same level, the total size of its “down buffers” at most equals the size of its “up buffer”. For example, Level $X^{3/2}$ in Figure 4 has one “up buffer” of size $\Theta(X^{3/2})$ and at most $X^{1/2}$ of “down buffers” each of size $\Theta(X)$. At most, the total size of the “down buffers” at level $X^{3/2}$ is $(X^{1/2} \times X = X^{3/2})$, which matches the size of the “up buffer” at level $X^{3/2}$. For any two consecutive levels, a “down buffer” at the larger level matches the size of the “up buffer” at the smaller level. For example, $X^{3/2}$ and X are consecutive levels as shown in Fig. 4, the size of a “down buffer” at level $X^{3/2}$ (i.e. the larger level) is X , which matches the size of the “up buffer” at level X (i.e. the smaller level) [34].

Generally, minimum and maximum items are corresponding to the highest priority item and the least priority item, respectively. At a level, items of its “up buffer” are disordered and their priorities are less than all items in the “down buffers” at that level. Items of a “down buffer” are disordered and their priorities are greater than items of the next “down buffer” at the same level. However, items of the “down buffers” at the very small levels are ordered and have the minimum item, i.e. the highest priority item. The priority queue algorithm knows the maximum item, which is corresponding to the least priority item, of each “down buffer” at all levels [31], [34], [29].

INSERT operation appends the new item i to the “up buffer” of the smallest level in the cache. Then the algorithm

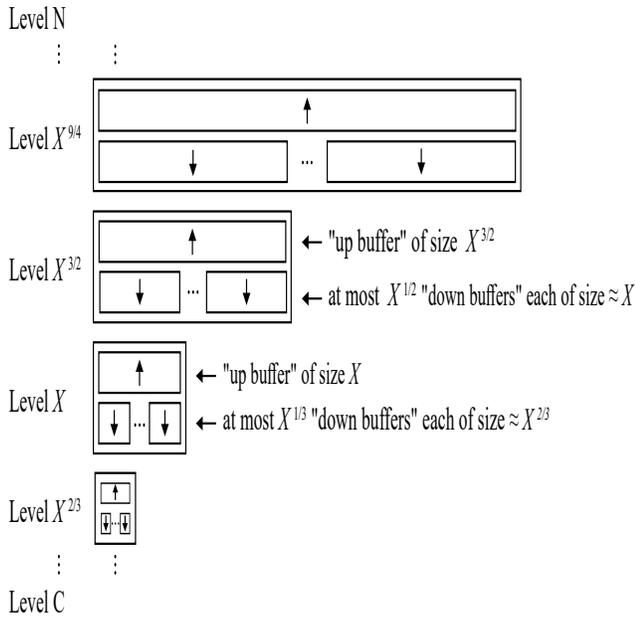


Fig. 4. Cache-Oblivious Priority Queue Levels.

locates i in an appropriate “down buffer” of the smallest level by comparing i to maximum items of the “down buffers” sequentially. If the selected “down buffer” has space, i is just added to it, otherwise, the “up buffer” swaps i with the maximum item in the selected “down buffer”. When the “up buffer” overflows, the algorithm performs the push operation[29].

To describe push operation, assume the “up buffer” of level X overflows, then the algorithm pushes X items of the level X “up buffer” into level $X^{3/2}$ “up buffer”. After that, the algorithm sorts X items and distributes them among the “down buffers” at level $X^{3/2}$ and possibly the “up buffer” at level $X^{3/2}$. Such that to allocate every item in its appropriate buffer, the algorithm scans the X items sequentially and visits “down buffers” at level $X^{3/2}$ in order. When the selected “down buffer” overflows, it is split in half to spawn a new “down buffer” at level $X^{3/2}$. However, if the number of allowed “down buffers” overflows, according to our example, the number of “down buffers” at level $X^{3/2}$ exceeds $X^{1/2}$, then the last “down buffer” at level $X^{3/2}$ is moved into the “up buffer” at level $X^{3/2}$. When this “up buffer” overflows, the algorithm pushes recursively $X^{3/2}$ items of the level $X^{3/2}$ “up buffer” into the level $X^{9/4}$ “up buffer” [29].

DELETE-MIN operation reverses the INSERT operation, therefore the algorithm deletes and pulls instead of inserts and pushes. The pull operation is a sort of reverse distribution step. Usually, “down buffers” of the smallest level are kept sorted in the cache and have the minimum item all the time, so the highest priority item is touched with zero memory transfers. However, there is no need to sort items in “down buffers” of the larger levels, we just need to keep track of the maximum items for the “down buffers” of larger levels. The smallest level in external memory is called the key level, where the algorithm consumes memory transfers to touch any item there.

In the cache-oblivious priority queue, most memory transfers are consumed to sort items. Consequently, the cache-oblivious priority queue provides INSERT and DELETE-MIN operations for one element with the cost of sorting, as shown in Equation 4 [29].

$$T(N) = O\left(\frac{1}{B} \log_{\frac{M}{B}} \frac{N}{B}\right) \quad (4)$$

F. Matrix Multiplication Algorithm

Matrix multiplication is the most essential matrix operation since it has significant applications in various fields. Examples are cryptography, wireless communication, computer graphics, computations in linear algebra, solution of linear systems of equations, the transformation of coordinate systems, and computational modeling [35], [36], [37], [38], [39].

Multiplying two matrices of size $N \times N$ using the standard matrix-multiplication algorithm incurs $O\left(\frac{N^3}{B}\right)$ memory transfers. However, the cache-oblivious algorithm uses the divide and conquer approach to solve matrix-multiplication problems. For simplicity, assume that A, B, and C are square matrices of size $N \times N$ for each. The cache-oblivious algorithm recursively partitions these matrices into quadrants, as shown in Fig. 5 [40], [41], [42].

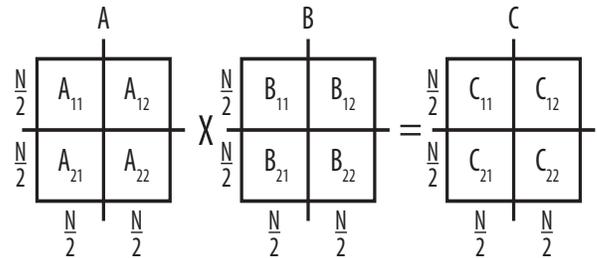


Fig. 5. Cache-Oblivious Algorithm Recursively Divides Matrices into quadrants.

The algorithm performs eight recursive matrix multiplications to update the four quadrants of C. At some point in this recursion, we get a base case, that the three sub-matrices fit in a certain number of B 's, or for the best base case when the three sub-matrices fit in the cache M . For cache complexity analysis, the recursion stops at the best base case, when the three sub-matrices of size $c\sqrt{M} \times c\sqrt{M}$, such that, c is a constant due to dividing cache into three sub-matrices. Then, the three sub-matrices fit in the cache, and accessing any element of them is free [43].

Nonetheless, this recursion is dominated by its leaves, so its cache complexity is the total number of leaves times the cache misses per leaf. Accordingly, each leaf of this recursion incurs $O\left(\frac{M}{B}\right)$ memory transfers, as the total number of leaves is:

$$8^{\lg N - \lg c\sqrt{M}} = 8^{\lg \frac{N}{c\sqrt{M}}} = \left(\frac{N}{c\sqrt{M}}\right)^{\lg 8} = \frac{N^3}{cM^{\frac{3}{2}}}$$

So, the total number of memory transfers is $\frac{M}{B} \times \frac{N^3}{cM^{\frac{3}{2}}}$. Consequently, Equation 5 shows the cache complexity of the

cache-oblivious matrix-multiplication algorithm using recursive block matrices, which is the optimal cache complexity for matrix-multiplication problems [40], [41], [42].

$$T(N) = O\left(\frac{N^3}{B\sqrt{M}}\right) \quad (5)$$

V. EXTENSION: CACHE-OBLIVIOUS NEURAL NETWORKS

Generally, a neural network iteratively learns by datasets, such that a neural network passes the whole dataset several times to learn correctly, as the neural network gets better performance in terms of mean squared error and accuracy [44]. The number of passes through the entire dataset is called epochs. To speed up the learning process, a dataset of size N is divided into mini-batches as a minimum learning unit, where a neural network updates its parameters after passing each mini-batch. Consequently, every epoch consists of several mini-batches, each of them having a size between 1 and N . For brevity, mini-batches are commonly called batches [8], [45], [46].

During the learning process, the neural network must be kept in the cache, but mini-batches are loaded and kicked out as needed. For the cache-oblivious neural network, the dataset must lay out in contiguous segments of memory, in any order, so the neural network scans the dataset in the same order it is stored [6]. The mini-batch size interacts with other hyper-parameters and must be optimized at the end to find the optimal size. However, if the size of a neural network is P , then the mini-batch size equals a multiple of B , as the selected multiple of B is recommended to be $\leq (M - P)$ to avoid incurring a memory transfer within a learning unit. In other words, it is recommended to select the mini-batch size as a power of two, that does not exceed $(M - P)$.

Accordingly, if a neural network of size P , using a dataset of size N , and a number of epochs E , then the cache complexity of this problem is described in Equation 6. As the number of epochs increases to get better results, the cache complexity of a neural network for a large problem (i.e. $P + N > M$) increases by at most the same factor. However, the number of epochs does not incur any memory transfer for a small problem (i.e. $P + N \leq M$).

$$T(N) = \begin{cases} O(\lceil \frac{P}{B} \rceil + \lceil \frac{N}{B} \rceil), & \text{if } P + N \leq M \\ O(\lceil \frac{P}{B} \rceil + E \times \lceil \frac{N}{B} \rceil), & \text{if } P + N > M \end{cases} \quad (6)$$

Neural networks greedily access memories, consequently, if N does not lay out in contiguous segments of memory, the cache complexity of a neural network will expand by a factor of N . Another extreme case that expands the cache complexity of a neural network, is when the N element is larger than B , so accessing any of the N elements initiates a memory transfer, too.

A. Experimental Results

To examine cache-oblivious neural networks, we implemented a cache simulator based on Intel i7 CPU and a memory hierarchy as described in Table II. Our cache simulator using

Python 3.6 is concerned with the L1 cache, whose block size is $8KB$, consequently, $M = 8MB$ and $B = 8KB$. Additionally, the simulator uses a tall full associative cache and the LRU as a page replacement policy.

TABLE II. EXPERIMENTAL MEMORY HIERARCHY

Memory level	Size
L1 Cache	8MB
L2 Cache	1GB
Main memory (RAM)	8GB

Furthermore, a 6-layer-stacked auto-encoder (SAE) model is used to classify the MNIST dataset as being created according to the authors of [8]. The experimental 6-layer SAE of 1139710 parameters occupied a $4.35MB$ of M . We examined two factors, the dataset sizes $\{0.63, 1.59, 3.18, 6.36, 9.69, 13.02, 16.36\}$ in MB and the number of epochs $\{0, 10, 20, \dots, 100\}$. Table III shows the experimental results, which are the total cache misses per epoch using different dataset sizes.

The available cache space for a dataset is $3.65MB$ because the experimental SAE occupied $4.35MB$ of M (i.e. $8 - 4.35$). Thus, three dataset sizes fit the free space of M as the rest dataset sizes are larger than the free space of M . Accordingly, we split the results into two figures; Fig. 6 shows the cache misses of the experimental SAE using small dataset sizes (i.e. $P + N \leq M$), and Fig. 7 shows the cache misses of the experimental SAE using large dataset sizes (i.e. $P + N > M$). Epoch zero represents the initial step when the SAE model is loaded to M without the dataset. Therefore, all experiments using any dataset size have the same number of cache misses at zero epoch, which equals 557 cache misses i.e. $\lceil \frac{\text{modelsize}}{B} \rceil = \lceil \frac{4.35MB}{8KB} \rceil$, as illustrated in Table III.

Fig. 6 shows that increasing dataset size increases the cache misses. However, increasing the number of epochs does not perform any additional cache misses. Because the experimental SAE and the dataset fit M . Fig. 7 shows that increasing dataset size increases the cache misses too. Additionally, increasing the number of epochs increases the cache misses by the same factor. Because the experimental SAE and the dataset are larger than M , consequently, every dataset access performs a cache miss. For example, the dataset size of $6.36MB$ in addition to the model size of $4.35MB$ needs $10.71MB$, which is greater than M . Therefore, the cache misses based on Equation 6 are $557 + E \times \lceil \frac{6.36MB}{8KB} \rceil$, as shown in Table III. However, if the dataset size is 3.18 , then the total needed space is $(3.18MB + 4.35MB = 7.53MB)$. Thus, the problem fits M , and the cache misses based on Equation 6 are $557 + \lceil \frac{3.18MB}{8KB} \rceil$, which is independent of E , as illustrated in Table III.

VI. SUMMARY

Cache-oblivious algorithms utilize all levels of memory hierarchy efficiently, without knowing their parameters or even the existence of memory hierarchy levels. Thus, cache-oblivious algorithms support portability and better memory performance. Cache complexities of cache-oblivious algorithms are denoted by the cache parameters i.e., the cache size M and the block size B , even though M and B are unknown for cache-oblivious algorithms in reality.

TABLE III. TOTAL CACHE MISSES PER EPOCH FOR THE EXPERIMENTAL SAE

epochs dataset size	0	10	20	30	40	50	60	70	80	90	100
0.63 MB	557	638	638	638	638	638	638	638	638	638	638
1.59 MB	557	761	761	761	761	761	761	761	761	761	761
3.18 MB	557	965	965	965	965	965	965	965	965	965	965
6.36 MB	557	8707	16857	25007	33157	41307	49457	57607	65757	73907	82057
9.69 MB	557	12967	25377	37787	50197	62607	75017	87427	99837	112247	124657
13.02 MB	557	17237	33917	50597	67277	83957	100637	117317	133997	150677	167357
16.36 MB	557	21497	42437	63377	84317	105257	126197	147137	168077	189017	209957

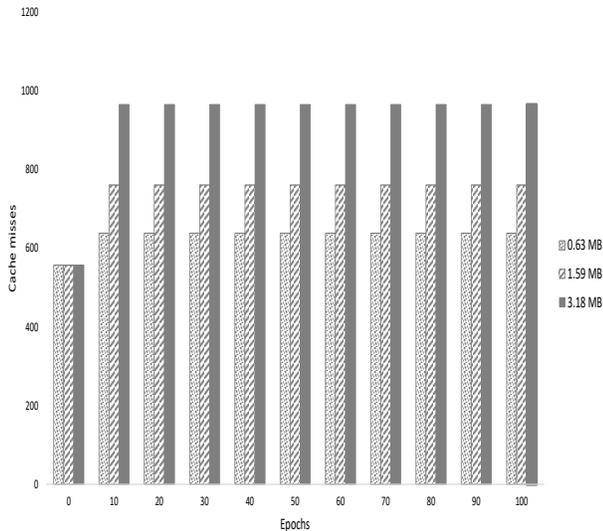


Fig. 6. Cache Misses of SAE using Small Datasets.

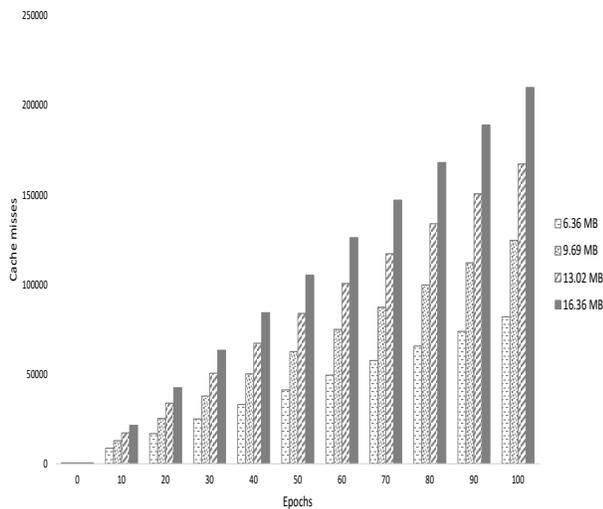


Fig. 7. Cache Misses of SAE using Large Datasets.

This paper discusses the cache complexities of the optimal recently known cache-oblivious algorithms for most essential problems, as summarized in Table IV. Notice that, the same input size N for all algorithms, but each of them incurs a

different number of cache misses according to the algorithm target. Generally, cache-oblivious algorithms utilize a divide and conquer approach with a base case based on the algorithm behavior. However, their cache complexities are calculated on different base cases that are proportional to either the block size B or the cache size M .

TABLE IV. CACHE-COMPLEXITY BASED ON CACHE-OBLIVIOUS MODEL

Cache-oblivious algorithm and data structure	Cache-complexity
Scanning	$\lceil \frac{N}{B} \rceil$
Binary search tree	$4 \log_B N$
$\frac{M}{B}$ -way merge sort	$\frac{N}{B} \log \frac{M}{B} \frac{N}{B}$
Priority queue	$\frac{1}{B} \log \frac{M}{B} \frac{N}{B}$
Recursive block-matrix multiplication	$\frac{N^3}{B\sqrt{M}}$
Small neural network ($P + N \leq M$)	$\lceil \frac{P}{B} \rceil + \lceil \frac{N}{B} \rceil$
Large neural network ($P + N > M$)	$\lceil \frac{P}{B} \rceil + E \times \lceil \frac{N}{B} \rceil$

Moreover, we introduce an extension that applies the cache-oblivious scanning approach to neural networks. In other words, to minimize the cache complexity of a neural network, the dataset must lay out in contiguous blocks of memory. When a neural network and its dataset are within cache size, the cache complexity is $\leq \frac{M}{B}$. Otherwise, cache complexity is growing by a factor that at most equals the number of epochs, as shown in Fig. 7. Nevertheless, if the dataset does not lay out in contiguous blocks of memory, then the cache complexity of a neural network expands, consequently, its learning process consumes unreasonable time.

REFERENCES

- [1] S. I. Serhan and H. M. Abdel-Haq, "Improving cache memory utilization," *World academy of science, engineering and technology*, vol. 26, pp. 299–304, 2007.
- [2] M. W. Ahmed and M. A. Shah, "Cache memory: An analysis on optimization techniques," *International Journal of Computer and IT*, vol. 4, no. 2, pp. 414–418, 2015.
- [3] L. Arge, G. S. Brodal, and R. Fagerberg, "Cache-oblivious data structures," pp. 545–565, 2018.
- [4] R. Sawant, B. H. Ramaprasad, S. Govindwar, and N. Mothe, "Memory hierarchies-basic design and optimization techniques," 2010.
- [5] N. Rahman, "Algorithms for hardware caches and tlb," pp. 171–192, 2003.
- [6] E. D. Demaine, "Cache-oblivious algorithms and data structures," *Lecture Notes from the EEF Summer School on Massive Data Sets*, vol. 8, no. 4, pp. 1–249, 2002.

- [7] R. E. Ladner, R. Fortna, and B.-H. Nguyen, "A comparison of cache aware and cache oblivious static search trees using program instrumentation," pp. 78–92, 2002.
- [8] I. Abuqaddom, B. A. Mahafzah, and H. Faris, "Oriented stochastic loss descent algorithm to train very deep multi-layer neural networks without vanishing gradients," *Knowledge-Based Systems*, vol. 230, p. 107391, 2021.
- [9] N. Ghatasheh, H. Faris, I. AlTaharwa, Y. Harb, and A. Harb, "Business analytics in telemarketing: cost-sensitive analysis of bank campaigns using artificial neural networks," *Applied Sciences*, vol. 10, no. 7, p. 2581, 2020.
- [10] E. D. Demaine, "Cache-oblivious algorithms and data structures," *Lecture Notes from the EEF Summer School on Massive Data Sets*, vol. 8, no. 4, pp. 1–249, 2002.
- [11] M. A. Bender, R. A. Chowdhury, R. Das, R. Johnson, W. Kuszmaul, A. Lincoln, Q. C. Liu, J. Lynch, and H. Xu, "Closing the gap between cache-oblivious and cache-adaptive analysis," in *Proceedings of the 32nd ACM Symposium on Parallelism in Algorithms and Architectures*, 2020, pp. 63–73.
- [12] E. Demaine and A. Schulz, "Mit 6.851 advanced data structures," *Angew. Chem. Int. Edit. Engl.*, vol. 6, pp. 53–67, 2010.
- [13] A. Lars, M. A. Bender, E. Demaine, C. Leiserson, and K. Mehlhorn, "Cache-oblivious and cache-aware algorithms," in *Cache-Oblivious and Cache-Aware Algorithms (Dagstuhl Seminar 04301)*. Schloss Dagstuhl, 2005.
- [14] J.-I. Agulleiro and J.-J. Fernandez, "Tuning the cache memory usage in tomographic reconstruction on standard computers with advanced vector extensions (avx)," *Data in brief*, vol. 3, pp. 16–20, 2015.
- [15] R. A. Chowdhury, "Algorithms and data structures for cache-efficient computation: theory and experimental evaluation," Ph.D. dissertation, 2007.
- [16] M. Frigo, C. E. Leiserson, H. Prokop, and S. Ramachandran, "Cache-oblivious algorithms," in *Foundations of Computer Science, 1999. 40th Annual Symposium on*. IEEE, 1999, pp. 285–297.
- [17] Y. Tang and W. Gao, "Processor-aware cache-oblivious algorithms*," in *50th International Conference on Parallel Processing*, 2021, pp. 1–10.
- [18] G. S. Brodal, R. Fagerberg, U. Meyer, and N. Zeh, "Cache-oblivious data structures and algorithms for undirected breadth-first search and shortest paths," in *Scandinavian Workshop on Algorithm Theory*. Springer, 2004, pp. 480–492.
- [19] A. Aggarwal, J. Vitter *et al.*, "The input/output complexity of sorting and related problems," *Communications of the ACM*, vol. 31, no. 9, pp. 1116–1127, 1988.
- [20] T. H. Chan, Y. Guo, W.-K. Lin, and E. Shi, "Cache-oblivious and data-oblivious sorting and applications," in *Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms*. SIAM, 2018, pp. 2201–2220.
- [21] M.-K. Lee, P. Michaud, J. S. Sim, and D. Nyang, "A simple proof of optimality for the min cache replacement policy," *Information Processing Letters*, vol. 116, no. 2, pp. 168–170, 2016.
- [22] M. A. Bender, E. D. Demaine, and M. Farach-Colton, "Cache-oblivious b-trees," in *Foundations of Computer Science, 2000. Proceedings. 41st Annual Symposium on*. IEEE, 2000, pp. 399–409.
- [23] A. Al-Adwan, R. Zaghoul, B. A. Mahafzah, and A. Sharieh, "Parallel quicksort algorithm on otis hyper hexa-cell optoelectronic architecture," *Journal of Parallel and Distributed Computing*, vol. 141, pp. 61–73, 2020.
- [24] S. W. A.-H. Baddar and B. A. Mahafzah, "Bitonic sort on a chained-cubic tree interconnection network," *Journal of Parallel and Distributed Computing*, vol. 74, no. 1, pp. 1744–1761, 2014.
- [25] B. A. Mahafzah, "Performance assessment of multithreaded quicksort algorithm on simultaneous multithreaded architecture," *The Journal of Supercomputing*, vol. 66, no. 1, pp. 339–363, 2013.
- [26] B. Elshqeir, M. Altarawneh, and A. Aloqaily, "Enhanced insertion sort by threshold swapping," *International Journal of Advanced Computer Science and Applications*, vol. 11, no. 6, 2020.
- [27] G. S. Brodal, "Cache-oblivious algorithms and data structures," in *Scandinavian Workshop on Algorithm Theory*. Springer, 2004, pp. 3–13.
- [28] G. S. Brodal, R. Fagerberg, and K. Vinther, "Engineering a cache-oblivious sorting algorithm," *Journal of Experimental Algorithmics (JEA)*, vol. 12, pp. 2–2, 2008.
- [29] G. S. Brodal, E. D. Demaine, J. T. Fineman, J. Iacono, S. Langerman, and J. I. Munro, "Cache-oblivious dynamic dictionaries with update/query tradeoffs," in *Proceedings of the twenty-first annual ACM-SIAM symposium on Discrete Algorithms*. Society for Industrial and Applied Mathematics, 2010, pp. 1448–1456.
- [30] M. Chrobak, M. Golin, J. I. Munro, and N. E. Young, "A simple algorithm for optimal search trees with two-way comparisons," *ACM Transactions on Algorithms (TALG)*, vol. 18, no. 1, pp. 1–11, 2021.
- [31] J. Iacono, R. Jacob, and K. Tsakalidis, "Cache-oblivious priority queues with decrease-key and applications to graph algorithms," *arXiv preprint arXiv:1903.03147*, 2019.
- [32] E. Shi, "Path oblivious heap: Optimal and practical oblivious priority queue," in *2020 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2020, pp. 842–858.
- [33] T. A. Assegie and H. D. Bizuneh, "Improving network performance with an integrated priority queue and weighted fair queue scheduling," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 19, no. 1, pp. 241–247, 2020.
- [34] J. Iacono, R. Jacob, and K. Tsakalidis, "External memory priority queues with decrease-key and applications to graph algorithms," in *27th Annual European Symposium on Algorithms (ESA 2019)*, vol. 144, 2019.
- [35] M. H. Qasem, A. A. Sarhan, R. Qaddoura, and B. A. Mahafzah, "Matrix multiplication of big data using mapreduce: a review," in *2017 2nd International Conference on the Applications of Information Technology in Developing Renewable Energy Processes & Systems (IT-DREPS)*. IEEE, 2017, pp. 1–6.
- [36] A. Sleit and A. Abusitta, "A visual cryptography based watermark technology for individual and group images," *Systems, Cybernetics and Informatics*, vol. 5, no. 2, pp. 24–32, 2008.
- [37] H. Faris, A. A. Heidari, A.-Z. Ala'M, M. Mafarja, I. Aljarah, M. Eshtay, and S. Mirjalili, "Time-varying hierarchical chains of salps with random weight networks for feature selection," *Expert Systems with Applications*, vol. 140, p. 112898, 2020.
- [38] J. E. H. Ali, E. Feki, and A. Mami, "Dynamic matrix control dmc using the tuning procedure based on first order plus dead time for infant-incubator," *International Journal of Advanced Computer Science and Applications*, vol. 10, no. 6, 2019.
- [39] I. Abuqaddom and A. Hudaib, "Cost-sensitive learner on hybrid smote-ensemble approach to predict software defects," in *Proceedings of the Computational Methods in Systems and Software*. Springer, 2018, pp. 12–21.
- [40] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein, "Introduction to algorithms," pp. 1–92, 2009.
- [41] Y. Tang, "Balanced partitioning of several cache-oblivious algorithms," in *Proceedings of the 32nd ACM Symposium on Parallelism in Algorithms and Architectures*, 2020, pp. 575–577.
- [42] M. Dusefante and R. Jacob, "Cache oblivious sparse matrix multiplication," in *Latin American Symposium on Theoretical Informatics*. Springer, 2018, pp. 437–447.
- [43] M. De Berg and S. Thite, "Cache-oblivious selection in sorted $x+y$ matrices," *Information processing letters*, vol. 109, no. 2, pp. 87–92, 2008.
- [44] P. M. Radiuk, "Impact of training set batch size on the performance of convolutional neural networks for diverse datasets," *Information Technology and Management Science*, vol. 20, no. 1, pp. 20–24, 2017.
- [45] O. A. Alrusaini, "Covid-19 detection from x-ray images using convoluted neural networks: A literature review," *International Journal of Advanced Computer Science and Applications*, vol. 13, no. 3, 2022.
- [46] H. Faris, S. Mirjalili, and I. Aljarah, "Automatic selection of hidden neurons and weights in neural networks using grey wolf optimizer based on a hybrid encoding scheme," *International Journal of Machine Learning and Cybernetics*, vol. 10, no. 10, pp. 2901–2920, 2019.

OvSbChain: An Enhanced Snowball Chain Approach for Detecting Overlapping Communities in Social Graphs

Jayati Gulati, Muhammad Abulaish
Department of Computer Science
South Asian University
New Delhi, India

Sajid Yousuf Bhat
Department of Computer Sciences
University of Kashmir
J&K, India

Abstract—Overlapping Snowball Chain is an extension to Snowball Chain, which is based on the concept of community formation in line to the snowball chaining process. The inspiration behind this approach is from the snowball sampling process, wherein a snowball grows to form chain of nodes, leading to the formation of mutually exclusive communities in Snowball Chain. In the current work, the nodes are allowed to be shared among different snowball chains in a graph, leading to the formation of overlapping communities. Unlike its predecessor Snowball Chain, the proposed technique does not require the use of any hyper-parameter which is often difficult to tune for most of the existing methods. The proposed algorithm works in two phases, where overlapping chains are formed in the first phase, and then they are combined using a similarity-based criteria in the second phase. The communities identified at the end of the second phase are evaluated using different measures, including *modularity*, *overlapping NMI* and *running time* over both real-world and synthetic benchmark datasets. The proposed Overlapping Snowball Chain method is also compared with eleven state-of-the-art community detection methods.

Keywords—Clustering coefficient; community detection; overlapping communities; snowball sampling; social graph

I. INTRODUCTION

In recent years, there has been a tremendous growth in the study of linked data in the form of networks, such as Internet, World Wide Web, and social networks. The relationships among the entities existing in these networks provide rich insights pertaining to various dynamic interactions and might prove to be beneficial in various applications [1]. To analyse and study these networks, graph is used as a data structure, which consists of a set of nodes joined by links or edges that can be labelled/unlabelled, directed/undirected, or signed/unsigned. The representation of an online social network is termed as *social graph*, which provides a good visualization and eases the interpretation of the network.

One of the emerging research areas in social network analysis is community detection, which digs deep into the social graph and mines the most dense subgraphs that are highly cohesive in nature. A community in a network is represented by a set of nodes with high density links among themselves, but low-density links among inter-community connections [2]. These subgraphs are called communities or modules. Community detection in a social graph mainly involves splitting it into its constituent functional groups. The task has

largely been addressed in a distinct community context wherein the communities are considered to be mutually exclusive. However, in case of real-world networks, community structures can be overlapping wherein a node belongs to multiple communities. A density-based approach called CMiner in [3], aims to find similarity among nodes and defines a distance function. Overlapping communities are identified based on this distance function. Another work in [4], detects overlapping communities along with their evolution, called as OCTracker. A similar work in [5], identifies hierarchical communities called HOCTracker which works for dynamic social networks.

The work in this paper aims to address this issue by proposing a novel overlapping community detection algorithm which extends the existing *SbChain* algorithm. The proposed method, named *OvSbChain*, starts with identification of the seed or core nodes in a social graph based on a node parameter, called *normalized degree*. The nodes in the entire social graph are ranked on this parameter and processed in a non-increasing order of their ranks. The method works in two phases. In the first phase, every node is paired with its best suited neighbor in accordance to a score value in each iteration. After several iterations, chains of nodes are formed that may share nodes with each other, i.e., there could be overlapping nodes among different chains. Therefore, the proposed technique is called overlapping snowball chains. The second phase tries to combine chains based on a similarity criteria as discussed in Section III, which finally leads to the formation of overlapping communities. Therefore, the technique focuses on resolving the problem in hand, i.e., community detection using an uncomplicated and elementary strategy. The major enhancements in this work can be summarized as follows:

- 1) *OvSbChain* introduces overlapping communities unlike *SbChain*, which produces only crisp communities.
- 2) There is no hyper-parameter tuning required in *OvSbChain*, hence, it always produces the same set of communities every time it is run.
- 3) *SbChain* uses a maximum common neighbor criteria for finding its best neighbor. Whereas, *OvSbChain* uses normalized degree function to find its best neighbor. Also, both the techniques differ in the way they find the seed nodes. This is discussed in detail in Section III.
- 4) The results are evaluated and compared based on

Nicosia modularity measure [6], two types of *ONMI* [7], [8] and their *running time*, as discussed in Section IV.

The proposed *OvSbChain* method is compared with eleven state-of-the-art community detection methods, including CFinder [9], LAIS [10], CONGA [11], PEACOCK [11], COPRA [12], SLPA [13], Demon [14], BIGCLAM [15], MULTICOM [16], Lemon [17] and ANGEL [18]. The results from all these methods are evaluated on different parameters including *modularity*, *ONMI* and *running time*, as discussed in Section IV.

The rest of the paper is organized as follows. Section II presents a brief review of the existing literatures on overlapping community detection. Section III presents the preliminary concepts, along with the proposed approach. This section also presents the functional details of the *OvSbChain* method. Section IV describes the details about the datasets, evaluation parameters, experimental settings, and analysis of the results. Section V concludes the paper and finally, Section VI provides future directions of research.

II. RELATED WORK

This section presents a brief description of the state-of-the-art in the area of overlapping community detection. A review of the current community detection methods is described in [19]. It segregates the detection methods into probability-based and deep learning-based. The classical methods use probability-based models for community identification. Whereas, complex networks are generally converted to lower dimensional data using deep learning methods so as to ease the process. A few other works like [20], [21], [22], discuss various community detection algorithms based on their weakness and strengths, performance of algorithm and other domains. We mainly discuss all the traditional approaches for overlapping community detection and compare them with *OvSbChain* in Section IV.

CFinder is an overlapping community detection technique that makes use of the Clique Percolation Method (CPM) [23] to identify the k -cliques in a network. A k -clique is a complete subgraph consisting of k nodes. This method finds dense groups of overlapping nodes in a network [9].

LAIS [10] is an algorithm that combines two functions List Aggregate (LA) and Improved Iterative Scan (IS^2). The LA procedure initializes the clusters, and the IS^2 procedure improves upon these set of clusters in an iterative manner. The IS procedure starts with a seed node and processes clusters by expanding or shrinking them according to a metric value, and IS^2 improves upon this by focussing on nodes within a cluster and its neighboring nodes, instead of considering the entire graph. The overall algorithm detects overlapping community in a network.

CONGA (Cluster-Overlap Newman Girvan Algorithm) [11] is an overlapping community detection algorithm that uses the concept of split-betweenness, i.e., it counts the shortest paths that exist between all pairs of nodes in the network. It keeps removing edges with high betweenness, and thus, keeps splitting the network into singleton clusters. The partition with the desired number of clusters is picked up. However, it requires number of communities as an input for the algorithm.

PEACOCK algorithm [11] consists of two phases; the first phase is similar to CONGA, where the network is split using split betweenness. The altered network is processed by a disjoint community detection algorithm, called centrality of detecting communities based on node centrality or CNM.

COPRA [12] technique extends the previous work on the label propagation by Raghavan, Albert, and Kumara [24], and it is able to detect overlapping communities in a social network. The main extension is to make the label and propagation step to include information about more than one community. Therefore, it allows each node to belong to up to v communities, where v is a hyper-parameter.

In SLPA (Speaker-listener Label Propagation algorithm) [13], the nodes store multiple labels, and act either as the provider or consumer of information. A node keeps gathering information about the observed labels without removing the previously stored label. The frequency of observation of a label by a node is directly related to the spreading of the label among other nodes. It requires a threshold input parameter that gives the minimum probability of occurrence of a label, before it is deleted from the memory of the node.

Demon (Democratic Estimate of the Modular Organization of a Network) [14] is a simple approach for community detection which works on the modular structure of networks. Firstly, each node finds and votes the communities present in its local neighborhood, using a label propagation algorithm. These local communities are merged to form a global collection by combining all the votes, leading to the formation of overlapping modules. However, this algorithm requires a minimum threshold parameter.

BIGCLAM (Cluster Affiliation Model for Big Networks) [15] is a model-based community detection algorithm that allows for identification of dense overlapping, hierarchical communities in massive networks. Each node-community pair is assigned a non-negative latent factor that decides the degree of membership between them. The probability of a connection between a pair of nodes in the network is modeled as a function of the shared community affiliations. Further, the communities are identified using non-negative matrix factorization methods and block stochastic gradient descent.

MULTICOM is another community detection technique that produces overlapping communities starting with an initial seed set. Local community is detected around the seed nodes using a transformation function. After this step, each node belongs to a single community. Thereafter, the transformation function is used to transform a node into its respective vector, that is clustered using a local clustering technique. For each cluster produced in the previous step, a ratio value is calculated using a function mentioned in [16]. The clusters having ratio value less than a pre-defined threshold are considered for further exploration. The process keeps repeating until the number of communities is greater than the set value or if there is no new seed.

In [17], the technique called Lemon (Local Expansion via Minimum One Norm) detects overlapping communities by finding a sparse vector in the local spectra span, such that all the seeds are in its support. The span of vector dimensions produced by random walk is used as an approximate invariant subspace, called the local spectra. However, this local spectral

approach is used for community detection from a small seed set.

ANGEL [18] is a faster successor of Demon that uses a bottom-up approach to find overlapping communities. It works in two phases, where the first phase produces local communities using ego network of the nodes. The second phase merges communities until convergence or a threshold value is met.

The work in [25], develops a PageRank algorithm with constraints so as to obtain tightly packed overlapping communities. Using probability-based methods, a walker avoids irrelevant communities. Therefore, it results in communities with good fitness score. In [26], a method called Adjacency Propagation Algorithm (APA) is developed using adjacent nodes as seed nodes. It uses a threshold parameter to identify subgraphs based on their intraconnectivity. Another work in [27], can produce disjoint as well as overlapping communities in a two-step process that uses genetic algorithm. In the first step, mean path length of a community is calculated in relation with its respective ER random graph. And the second step shrinks the search space by selecting a subset of nodes. Another work in [28], influential nodes are identified to form local communities. These communities expand as nodes join these local communities. Overlapping communities are merged and evaluated on a model.

An application-based work in [29], exploits community detection to protect the privacy of individuals on social platforms. It discusses community detection attacks and rewiring of connections for development of effective attack approach.

OvSbChain approach focuses on local community detection using graph parameters such as degree and global clustering coefficient. If these local communities are identical they are merged. The motivation behind this work is that it exploits simple topological features of the graph to detect communities without any expensive overhead in two simple levels, (i) formation of local communities, and (ii) combining local communities based on two criteria.

III. PROPOSED APPROACH

The *OvSbChain* approach discussed in this section is an extension to the previously developed *SbChain* [30] method. It detects overlapping communities, i.e., nodes are allowed to be shared among more than one community. The approach works on two levels. In the first level, it starts with finding the best suited pairs of nodes according to an initial criterion. This level ends up with formation of overlapping snowball chains. In the second level, these chains are merged to form the larger chains, and eventually form communities based on global clustering coefficient or majority overlapping criteria.

A. Preliminaries

For a graph $G(V, E)$, V represents the set of vertices or nodes in the graph, i.e. $\{v \in V\}$, where n is the number of nodes. And E is the set of edges, i.e., $\{e_{uv} = (u, v) : u, v \in V\}$. This section presents the details about frequently used terms and their meanings, as mentioned in table I.

OvSbChain works at two levels that are described in the following paragraphs:

TABLE I. NOTATIONS AND THEIR DESCRIPTIONS

Notation	Description
$\mathcal{N}(v)$	Set of immediate neighbors of a node v
$k(v) = \mathcal{N}(v) $	Degree of a node v
k_{max}	Maximum degree value in the graph
$\mathcal{N}_{best}(v)$	Best scoring neighbor of node v
$s^{(n)}$	n^{th} snowball chain
$GCC(s^{(n)})$	Global clustering coefficient of a snowball chain $s^{(n)}$

1) *Level-I*: It starts by finding the seed nodes and sorting them in non-increasing order, based on the following criteria so as to begin the processing.

- 1) Seed function - A seed v can be identified by sorting nodes according to their normalized degree value function, given by equation 1. This also represents the score $score(v)$ of a node v .

$$score(v) = k(v)/k_{max} \quad (1)$$

These sorted nodes are processed in non-increasing order of this function value. It should be noted that *SbChain* used a combination of normalized degree and normalized local clustering coefficient for sorting of nodes.

- 2) $\mathcal{N}_{best}(v)$ function - The best suited neighbor for a seed v is identified using the same score value, i.e., the normalized degree. This neighbor further combines with the seed v to form a snowball chain. Whereas, *SbChain* used maximum number of overlapping neighbors for finding its best neighbor.

It should be noted that these functions have been chosen and designed empirically.

2) *Level-II*: The second level starts with the chains formed in the first level. These chains are merged to form communities, so as to eliminate almost similar chains. The snowball pairs/chains formed in first level are combined based on global clustering coefficient (GCC) or majority overlapping criteria to form a community. GCC signifies the number of closed triangles to the number of triplets in a graph. Therefore, the technique focuses on finding higher values of GCC for a community, so as to find coherent communities. The first criteria involves calculation of GCC of the formed community, along with GCC of each individual snowball chain. If the combined GCC is higher than the GCC of each chain, then their combination is permitted, otherwise it is discarded, i.e., the chains remain undisturbed. Communities can also be combined as per the second criteria of majority overlapping. This allows communities to get merged if they have atleast 70% overlapping nodes. This percentage is decided empirically, as the value of communities do not change after this point. Also, the minimum percentage overlap was decided to be above 50% so as to form coherent communities. The majority overlapping test prevents the existence of two or more similar communities.

It should be noted that *OvSbChain* creates overlapping communities because it does not follow *non-redundant node strategy*, previously used by *SbChain*. According to this

Algorithm 1: bestNeighbor($v, \mathcal{N}(v)$)

Input : Node v , neighbor list $\mathcal{N}(v)$
Output: Best neighbor of v i.e., $\mathcal{N}_{best}(v)$

```

1  $maxWeight \leftarrow 0$ 
2 foreach  $v \in \mathcal{N}(v)$  do
3   if  $score(v) \geq maxWeight$  then
4      $maxWeight \leftarrow score(v)$ 
5      $\mathcal{N}_{best}(v) \leftarrow v$ 
6   end
7 end
8 return  $\mathcal{N}_{best}(v)$ 

```

strategy, a node could join with a single node per iteration which creates mutually exclusive communities. The focus of *OvSbChain* is to develop communities that share nodes among themselves. Therefore, it discards this strategy and allows a node to be a part of multiple chains within a single iteration itself.

TABLE II. DIFFERENT TYPES OF REAL-WORLD DATASETS

Dataset	Nodes	Edges
Zachary [31]	34	78
Dolphin [32]	62	159
Football [33]	115	613
Books ¹	105	441
Netscience [34]	379	914
Jazz [35]	198	5484
Email [36]	1133	5451
Power [37]	4941	6594
Blogs [38]	3982	6803
Protein [39]	2445	6265

TABLE III. PARAMETERS USED TO GENERATE LFR-1K NETWORK

Parameter	Value
Nodes (N)	1000
Average degree ($\langle k \rangle$)	15
Minimum community size (c_{min})	20
Maximum community size (c_{max})	50
Maximum degree (k_{max})	50
Number of overlapping nodes (o_n)	100
Number of memberships of the overlapping nodes (o_m)	30
Mixing parameter (μ)	[0.1, 0.5]

B. Algorithm

As discussed in the algorithm 2, *OvSbChain* starts with the pre-processing, i.e., it calculates neighbor list $\mathcal{N}(v)$, degree list $k(v)$ and $score(v)$ (equation 1) for each node v in the social graph. These nodes are then sorted in non-increasing order of their respective $score$ and processed one at a time. Snowball chains are formed by finding the best neighbor (algorithm 1) for each node on the basis of this $score$ value itself, i.e., for a given node v , the best neighboring node $\mathcal{N}_{best}(v)$ with highest value of $score$ parameter is chosen.

In the first iteration, best suited node pairs are combined. The snowball chains $s^{(n)}$ so formed grow internally and new chains are also formed in each iteration, as the nodes find their matches. This sums up the level-I of the proposed technique.

The level-II starts with calculation of global clustering coefficient $GCC(s^{(n)})$ for each snowball chain $s^{(n)}$ formed in level-I. These chains are combined and added to community list C if the GCC of the union of two chains $GCC(s^{(j)} \cup s^{(k)})$

Algorithm 2: OvSbChain(G)

Input : A graph $G(V, E)$
Output: Community list C

```

1 Pre-processing calculates  $\mathcal{N}(v)$ ,  $k(v)$ , and  $score(v)$  for each node  $v$ 
2 Arrange  $score(v)$  in non-increasing order
3 Initialize new lists snowball  $s$ , community  $C$ 
4  $i \leftarrow 0$ 
   // Level-I
5 foreach  $v \in score.keys$  do
6    $\mathcal{N}_{best}(v) \leftarrow bestNeighbor(v, \mathcal{N}(v))$ 
   // Algorithm 1
7   if  $s = \emptyset$  then
8      $i \leftarrow i + 1$ 
9     Append  $\langle v, \mathcal{N}_{best}(v) \rangle$  into  $s^{(i)}$ 
10    Goto 5
11  end
12   $counter \leftarrow 0$ 
13  for  $j \leftarrow 1$  to  $len(s)$  do
14    if  $\mathcal{N}_{best}(v) \in s^{(j)}$  and  $v \notin s^{(j)}$  then
15      Append  $\langle v \rangle$  into  $s^{(j)}$ 
16    else
17      if  $\mathcal{N}_{best}(v) \notin s^{(j)}$  and  $v \in s^{(j)}$  then
18        Append  $\langle \mathcal{N}_{best}(v) \rangle$  into  $s^{(j)}$ 
19      else
20         $counter \leftarrow counter + 1$ 
21      end
22    end
23  end
24  if  $counter = i$  then
25     $i \leftarrow i + 1$ 
26    Append  $\langle v, \mathcal{N}_{best}(v) \rangle$  into  $s^{(i)}$ 
27  end
28 end
   // Level-II
29 while  $C \neq s$  do
30   if  $C \neq \emptyset$  then
31      $s \leftarrow C$ 
32      $C \leftarrow \emptyset$ 
33   end
34   foreach  $s^{(j)} \in s$  do
35     foreach  $s^{(k)} \in s$  do
36       if  $(GCC(s^{(j)} \cup s^{(k)}) > GCC(s^{(j)})$  and
37          $GCC(s^{(j)} \cup s^{(k)}) > GCC(s^{(k)}))$  or
38          $\frac{||s^{(j)} \cap s^{(k)}||}{\min(||s^{(j)}||, ||s^{(k)}||)} > 0.7$  then
39         Append  $\langle s^{(j)}, s^{(k)} \rangle$  into  $C$ 
40       else
41         Append  $\langle s^{(j)} \rangle, \langle s^{(k)} \rangle$  into  $C$ 
42       end
43     end
44   end
45 return  $C$ 

```

is greater than either of their individual GCC or if majority of their nodes overlap (i.e., $\geq 70\%$) as mentioned in step 36 of the algorithm. The chains keep combining until both of the criteria fail. If the chains do not combine with other chains, they are directly added to C . The end result is the final set of communities C .

IV. EXPERIMENTAL SETUP AND RESULTS

In this section, the performance of the *OvSbChain* algorithm is evaluated over different datasets using various parameters. The *OvSbChain* is compared with several other overlapping community detection techniques. The following

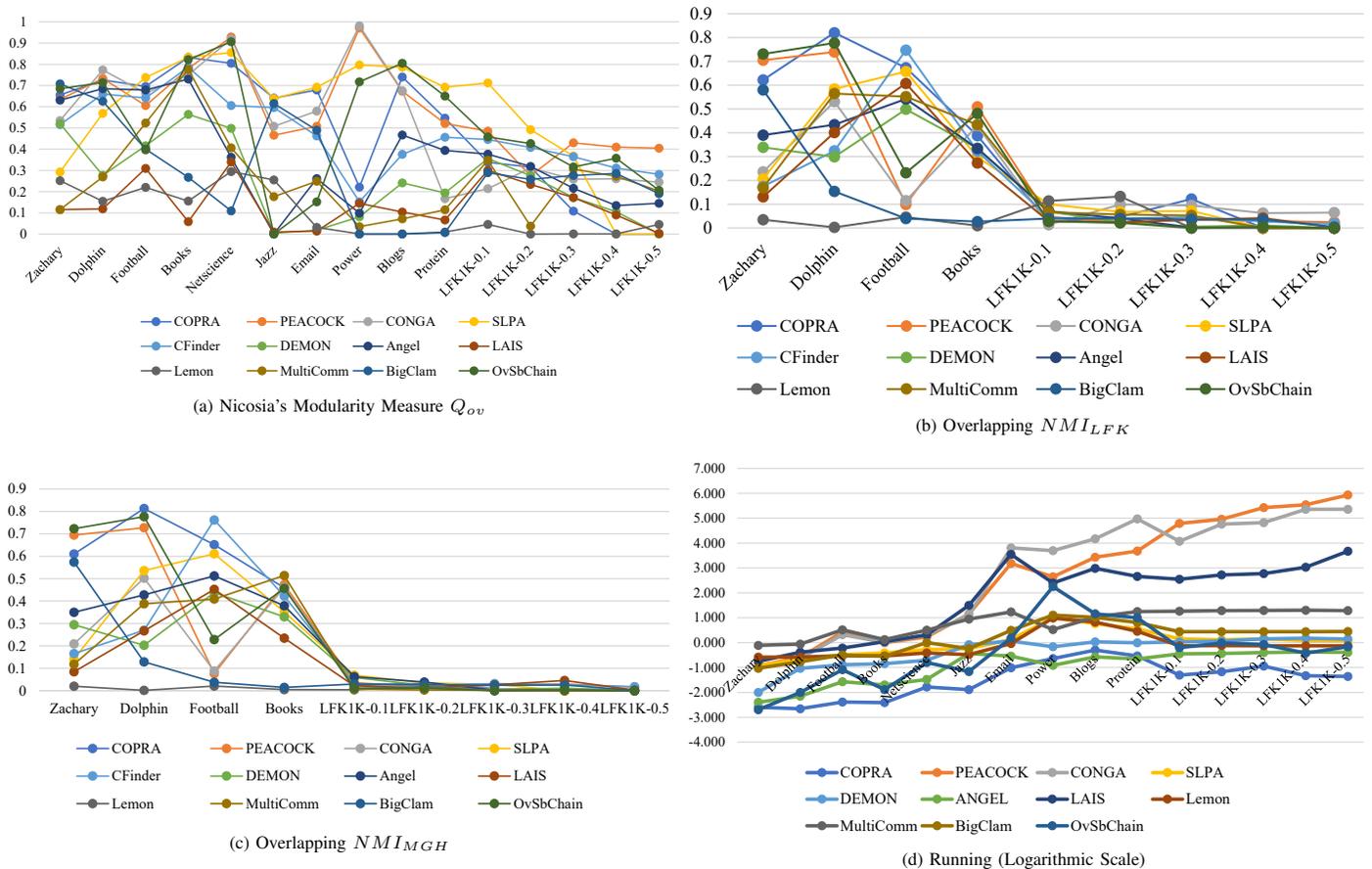


Fig. 1. Comparison of *OvSbChain* with Various Techniques on different Evaluation Metrics

subsections discuss the various datasets used in our experimental evaluations and all the parameters used for assessment of the identified communities.

A. Dataset

The efficacy of *OvSbChain* and other techniques is evaluated over ten real-world datasets and five computer-generated Lancichinetti Fortunato Radicchi (LFR) benchmark datasets [40], as discussed in Tables II and III. The LFR benchmark datasets consists of 1000 nodes with value of the mixing parameter (μ) varying from 0.1 to 0.5. Hence, the datasets are named as LFR1K-0.1, LFR1K-0.2, ..., LFR1K-0.5, respectively.

B. Evaluation Metrics

The communities identified as a result of algorithm 2 are analyzed by an overlapping *modularity* measure given by [6], two types of *ONMI* (Overlapping Normalized Mutual Information), and their *running time*.

It should be noted that the modularity measure given by [6] is represented as Q_{ov} . By definition, $Q_{ov} = 0$, for singleton communities or if all nodes belong to a community. Q_{ov} uses a belonging coefficient for each node which defines the percentage contribution of a node in a community. The sum of this coefficient in 1, for each node.

ONMI is an extension of the NMI score that accommodates overlapping partitions within a network. There are two types of ONMI used in this section; one is LFK (Lancichinetti Fortunato Kertesz) [7], which is referred as NMI_{LFK} , but it overestimates the similarity of two clusters in some cases. To fix this, another ONMI called MGH (McDaid Greene Hurley) is used. This version uses a different normalization than the original LFK based ONMI [8], and it is represented as NMI_{MGH} .

C. Results

Techniques like COPRA, PEACOCK, CONGA, SLPA, CFinder, Demon, and ANGEL use a parameter for tuning. Hence, the values represented in this paper are the best values for Q_{ov} . Fig. 1a shows the results of various overlapping techniques compared with *OvSbChain* on Q_{ov} , respectively. The same is also represented via Table IV. Also, Fig. 2a represents the number of datasets for each technique that have their respective value greater than or equal to 80% of the maximum Q_{ov} that exists for all the techniques. It can be observed that *OvSbChain* has an above average performance in terms of Q_{ov} . Though other techniques are seen to show a better value in terms of Q_{ov} , it is seen that the respective ONMI values drops. Hence, high modularity does not necessarily guarantee good partitions.

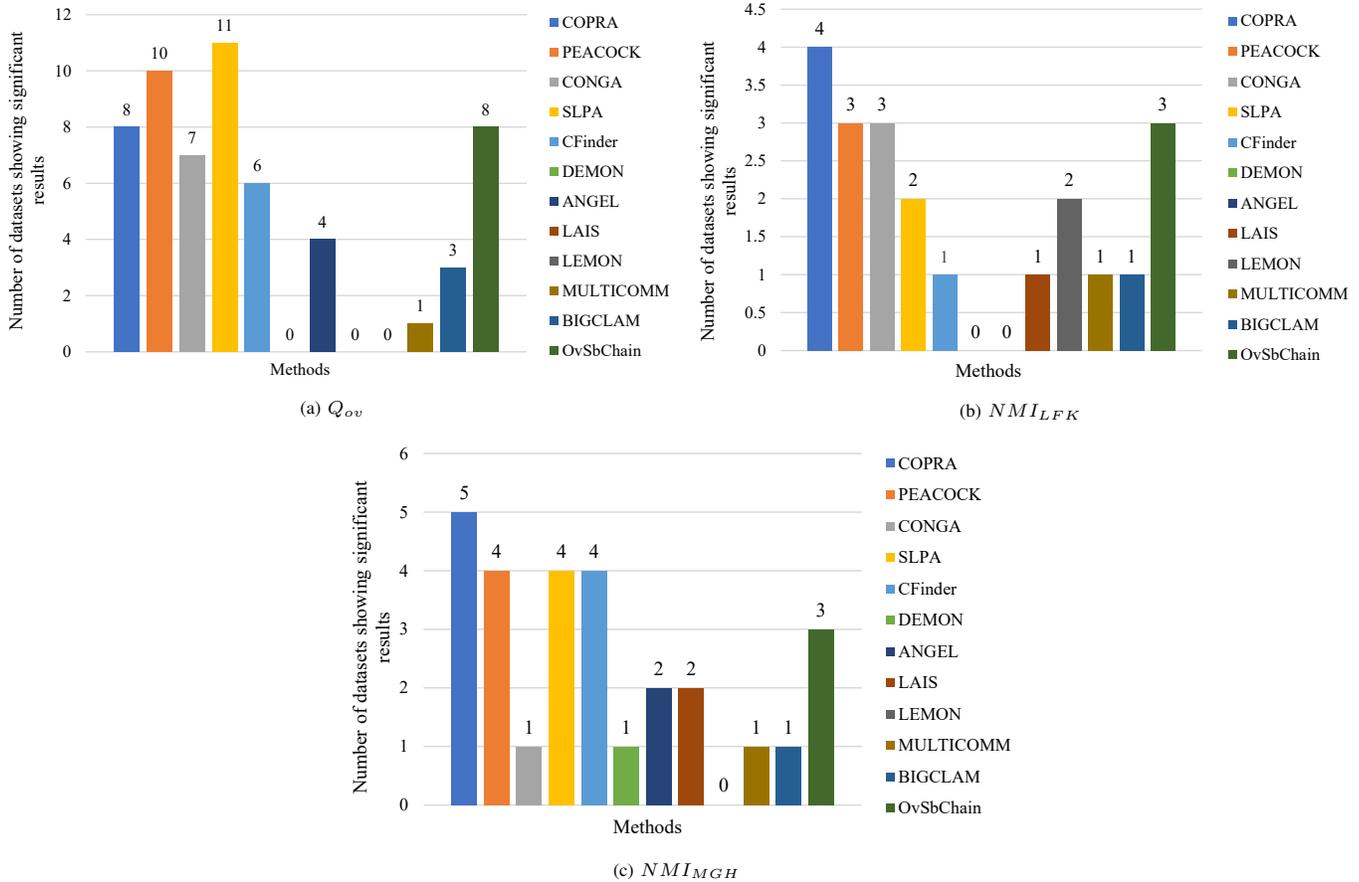


Fig. 2. Comparison of Various Techniques on the Number of Datasets having Values Greater than or Equal to 80% of the Maximum Existing Value of Different Evaluation Metrics

It can be seen that although modularity values are comparable or average in comparison to existing techniques, the ONMI values are promising. As an example, SLPA has the highest modularity among all techniques, it does not produce high ONMI values. *OvSbChain* is faster for smaller datasets and produces comparable or better results for certain cases in terms of both NMI_{LFK} and NMI_{MGH} .

Both NMI_{LFK} and NMI_{MGH} are calculated and compared on both real-world and LFR datasets, as shown in Fig. 1b and 1c for *OvSbChain* and other techniques. *OvSbChain* is seen to perform well in most of the cases. Fig. 2b and 2c also show the comparison of the number of datasets that have NMI values greater than or equal to 80% of the maximum existing value of NMI (among all the given datasets). Tables VI and VII show both the ONMI values. It can be observed that the performance of *OvSbChain* is above average for both NMI_{LFK} and NMI_{MGH} measures.

A comparison of the running time of all the techniques is presented in Fig. 1d. Logarithmic scale is used for this comparison because it provides a better visualization. CFinder technique is excluded from this comparison because it does not mention the time it takes to evaluate the communities so formed. It can be observed that *OvSbChain* works fast on smaller datasets, and it is comparable to other techniques on

larger datasets. The same can be seen through table V. As mentioned before, a few techniques use a parameter which needs to be defined every time they are executed. Therefore, in our experimental evaluation, these techniques were run for different parameter values and the best value for Q_{ov} was chosen and the corresponding ONMI and run time values are represented. On the other hand, our proposed *OvSbChain* approach does not need any parameter value to be set, hence, produces the same result every time it is run.

V. CONCLUSION

It can be seen that the technique *OvSbChain* discussed in the current article works well on real-world datasets with good results in terms of NMI_{LFK} and NMI_{MGH} . It gives comparable results on a few benchmark datasets as well. It should be noted that the running speed of the algorithm was at par with other techniques, or even better in a few cases. The experiments show average results on modularity measure as well. *OvSbChain* does not use any external parameter like most of its counterparts. Also, it produces the same results every time it is run, unlike the other techniques, e.g., COPRA. It gives different results each time it is run (for same parameter value). Hence, it is run for ten times, and the results are averaged. Therefore, it can be established that our technique works well without any parameter tuning, unlike the other

approaches. The overhead of calculations involved in the technique slows it down, but that can be resolved using better hardware options.

VI. FUTURE WORK

The future scope of improvement includes extension of the technique to directed graphs as real-world networks are generally directed in nature. OvSbChain can be improvised to find faster and high coverage seed nodes for snowball chains formation and eventually communities.

TABLE IV. COMPARISON AMONG DIFFERENT OVERLAPPING COMMUNITY DETECTION TECHNIQUES ON VARIOUS REAL-WORLD AND BENCHMARK DATASETS ON NICOSIA'S OVERLAPPING MODULARITY MEASURE

Dataset	Techniques											
	COPRA	Peacock	CONGA	SLPA	CFinder	Demon	ANGEL	LAIS	Lemon	MULTICOM	BIGCLAM	OvSbChain
Zachary	0.655 (5)	0.634 (2)	0.534 (3)	0.292 (0.2)	0.515 (3)	0.519 (0.5)	0.631 (0.4-0.6)	0.115	0.252	0.115	0.708	0.685
Dolphin	0.726 (4)	0.732 (2)	0.773 (4)	0.569 (0.4-0.5)	0.659 (3)	0.276 (0.4)	0.685(0.3)	0.119	0.154	0.269	0.625	0.714
Football	0.695 (2)	0.605 (2)	0.664 (2)	0.737 (0.3-0.5)	0.641 (4)	0.415 (0.6)	0.678 (0.3)	0.309	0.220	0.523	0.400	0.397
Books	0.832(3)	0.779 (2)	0.749 (3)	0.832 (0.5)	0.786 (3)	0.564 (0.6)	0.730 (0.6)	0.058	0.155	0.776	0.267	0.822
Netscience	0.804 (4)	0.929 (5)	0.919 (5)	0.855 (0.4-0.5)	0.605 (3)	0.497 (0.3)	0.362 (0.4)	0.339	0.294	0.406	0.109	0.905
Jazz	0.640 (3)	0.465 (3)	0.508 (5)	0.638 (0.3-0.5)	0.595 (10)	0.004 (0.8)	0.0004 (1)	0.009	0.255	0.176	0.615	0.000 ¹
Email	0.678 (2)	0.507 (2)	0.579 (5)	0.692 (0.5)	0.462 (3)	0.018 (0.5)	0.261 (0.6)	0.013	0.031	0.249	0.489	0.159
Power	0.221 (5)	0.970 (5)	0.981 (5)	0.797 (0.5)	0.152 (3)	0.081 (0.1)	0.099 (0.1-0.2)	0.145	0.0007	0.035	0.000 ¹	0.717
Blogs	0.740 (9)	0.672 (9)	0.675 (9)	0.786 (0.5)	0.376 (3)	0.241 (0.2)	0.466 (0.3)	0.104	0.001	0.071	0.000 ¹	0.804
Protein	0.545 (5)	0.520 (5)	0.167 (5)	0.692 (0.5)	0.456 (3)	0.194 (0.5)	0.393 (0.4)	0.066	0.009	0.114	0.007	0.650
LFR1K-0.1	0.337 (2)	0.484 (2)	0.214 (5)	0.712 (0.3)	0.446 (4)	0.355 (0.5)	0.377 (0.6)	0.300	0.045	0.347	0.289	0.458
LFR1K-0.2	0.315 (3)	0.275 (3)	0.311 (5)	0.492 (0.5)	0.408 (4)	0.275 (0.5)	0.319 (0.4)	0.234	0.0001	0.037	0.258	0.426
LFR1K-0.3	0.108 (2)	0.429 (2)	0.259 (5)	0.368 (0.5)	0.364 (4)	0.172 (0.5)	0.215 (0.5)	0.172	0.0006	0.305	0.275	0.316
LFR1K-0.4	0.000 ¹ (5)	0.409 (2)	0.260 (4)	0.000 ¹ (0.1)	0.310 (3)	0.105 (0.4)	0.134 (0.5)	0.089	0.001	0.272	0.285	0.357
LFR1K-0.5	0.000 ¹ (2)	0.404 (2)	0.244 (2)	0.000 ¹ (0.1)	0.281 (3)	0.001 (0.2)	0.145 (0.6)	0.003	0.045	0.202	0.190	0.206

TABLE V. COMPARISON AMONG DIFFERENT OVERLAPPING COMMUNITY DETECTION TECHNIQUES ON VARIOUS REAL-WORLD AND BENCHMARK DATASETS BASED ON (IN SECONDS)

Time	Dataset	Techniques										
		COPRA	Peacock	CONGA	SLPA	Demon	ANGEL	LAIS	Lemon	MULTICOM	BIGCLAM	OvSbChain
	Zachary	0.003	0.116	0.104	0.120	0.010	0.004	0.175	0.259	0.774	0.093	0.002
	Dolphin	0.002	0.254	0.229	0.208	0.092	0.007	0.413	0.262	0.873	0.171	0.010
	Football	0.004	2.668	2.081	0.321	0.129	0.027	0.607	0.293	3.2	0.315	0.082
	Books	0.004	0.979	1.007	0.385	0.143	0.020	1.092	0.288	1.316	0.281	0.013
	Netscience	0.017	1.684	1.957	0.525	0.194	0.033	2.049	0.381	3.1	0.998	0.158
	Jazz	0.013	14.065	13.1	0.567	0.830	0.362	31.2	0.328	8.8	0.575	0.069
	Email	0.096	1516.04	6428.1	1.352	1.148	0.285	3526.8	0.921	17.02	3.067	1.502
	Power	0.226	442.4	4948.8	9.807	0.680	0.122	246.9	9.7	3.3	12.7	177.1
	Blogs	0.498	2675.7	14746.6	5.756	1.066	0.266	971.2	6.6	10.691	10.3	14.3
	Protein	0.289	4781.4	93087.7	3.535	0.977	0.221	454.5	2.815	17.5	6.3	9.969
	LFR1K-0.1	0.049	61766.5	11830.8	1.408	1.068	0.353	353.5	0.741	18.1	2.7	0.627
	LFR1K-0.2	0.068	90091.7	57434.7	1.294	1.148	0.362	522.1	0.760	19.3	2.7	0.968
	LFR1K-0.3	0.114	265414.9	65968.1	1.409	1.415	0.399	590.1	0.748	19.6	2.7	0.833
	LFR1K-0.4	0.047	344527.9	228258.1	1.203	1.493	0.419	1069.1	0.747	20.1	2.731	0.375
	LFR1K-0.5	0.043	851668.6	229433.5	1.178	1.378	0.404	4712.3	0.737	19.3	2.7	0.682

TABLE VI. COMPARISON AMONG DIFFERENT OVERLAPPING COMMUNITY DETECTION TECHNIQUES ON VARIOUS REAL-WORLD AND BENCHMARK DATASETS ON NMI_{LFFK} MEASURES

NMI_{LFFK} Dataset	Techniques											
	Copra	Peacock	CONGA	SLPA	CFinder	Demon	ANGEL	LAIS	Lemon	MULTICOM	BIGCLAM	OvSbChain
Zachary	0.622	0.705	0.236	0.205	0.174	0.338	0.390	0.131	0.034	0.168	0.580	0.730
Dolphin	0.820	0.739	0.531	0.583	0.323	0.298	0.433	0.401	0.002	0.564	0.153	0.776
Football	0.672	0.098	0.115	0.657	0.747	0.498	0.541	0.607	0.046	0.552	0.039	0.231
Books	0.387	0.588	0.400	0.292	0.323	0.336	0.333	0.273	0.004	0.433	0.027	0.484
LFRIK-0.1	0.024	0.047	0.014	0.100	0.043	0.068	0.071	0.028	0.114	0.067	0.039	0.029
LFRIK-0.2	0.046	0.022	0.099	0.067	0.039	0.024	0.040	0.024	0.132	0.057	0.039	0.022
LFRIK-0.3	0.000	0.034	0.094	0.072	0.046	0.005	0.002	0.035	0.000	0.053	0.037	0.000
LFRIK-0.4	0.000	0.029	0.062	0.000	0.027	0.0104	0.002	0.041	0.000 ¹	0.000	0.035	0.004
LFRIK-0.5	0.000	0.024	0.065	0.000	0.0175	0.000	0.000	0.002	0.000	0.000	0.004	0.000

TABLE VII. COMPARISON AMONG DIFFERENT OVERLAPPING COMMUNITY DETECTION TECHNIQUES VARIOUS REAL-WORLD AND BENCHMARK DATASETS ON NMI_{MGH} MEASURES

NMI_{MGH} Dataset	Techniques											
	COPRA	Peacock	CONGA	SLPA	CFinder	Demon	ANGEL	LAIS	Lemon	MULTICOM	BIGCLAM	OvSbChain
Zachary	0.610	0.694	0.208	0.142	0.165	0.294	0.349	0.084	0.020	0.117	0.574	0.723
Dolphin	0.813	0.727	0.502	0.536	0.269	0.202	0.427	0.267	0.001	0.387	0.129	0.776
Football	0.651	0.076	0.088	0.610	0.762	0.437	0.512	0.452	0.020	0.409	0.0381	0.223
Books	0.459	0.473	0.460	0.351	0.421	0.330	0.379	0.234	0.004	0.514	0.014	0.456
LFRIK-0.1	0.022	0.035	0.008	0.069	0.029	0.0586	0.0628	0.022	0.0042	0.008	0.029	0.011
LFRIK-0.2	0.033	0.017	0.021	0.034	0.024	0.023	0.0389	0.0122	0.0188	0.003	0.027	0.010
LFRIK-0.3	0.009	0.023	0.008	0.028	0.031	0.0054	0.0022	0.026	0.000	0.002	0.024	0.000
LFRIK-0.4	0.000	0.029	0.008	0.000	0.0277	0.010	0.0021	0.046	0.000	0.000	0.025	0.003
LFRIK-0.5	0.000	0.016	0.005	0.000	0.017	0.000	0.000	0.003	0.000	0.000	0.001	0.000

REFERENCES

- [1] M. E. J. Newman, "The structure and function of complex networks," *Society for Industrial and Applied Mathematics (SIAM) Review*, vol. 45, no. 2, pp. 167–256, March 2003.
- [2] A. Lancichinetti and S. Fortunato, "Community detection algorithms: A comparative analysis," *Physical Review E*, vol. 80, no. 5, pp. 056 117–(1–11), November 2009.
- [3] S. Y. Bhat and M. Abulaish, "A density-based approach for mining overlapping communities from social network interactions," in *Proceedings of the 2nd International Conference on Web Intelligence, Mining and Semantics (WIMS), Craiova, Romania*, June 2012, pp. 1–7.
- [4] —, "OCTracker: A density-based framework for tracking the evolution of overlapping communities in OSNs," in *Proceedings of IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), Istanbul, Turkey*, August 2012, pp. 501–505.
- [5] —, "HOCTracker: Tracking the evolution of hierarchical and overlapping communities in dynamic social networks," *IEEE Transactions on Knowledge and Data Engineering*, vol. 27, no. 4, pp. 1019–1031, April 2015.
- [6] V. Nicosia, G. Mangioni, V. Carchiolo, and M. Malgeri, "Extending the definition of modularity to directed graphs with overlapping communities," *Journal of Statistical Mechanics Theory and Experiment*, vol. 2009, pp. P03 024–P03 046, February 2008.
- [7] A. Lancichinetti, S. Fortunato, and J. Kertész, "Detecting the overlapping and hierarchical community structure in complex networks," *New Journal of Physics*, vol. 11, no. 3, pp. 033 015–033 032, March 2009.
- [8] A. McDaid, D. Greene, and N. Hurley, "Normalized mutual information to evaluate overlapping community finding algorithms," *Computing Research Repository*, October 2011.
- [9] B. Adamcsek, G. Palla, I. J. Farkas, I. Derényi, and T. Vicsek, "CFinder: Locating cliques and overlapping modules in biological networks," *Bioinformatics*, vol. 22, no. 8, pp. 1021–1023, April 2006.
- [10] J. Baumes, M. Goldberg, and M. MagdonIsmail, "Efficient identification of overlapping communities," in *Proceedings of International Conference on Intelligence and Security Informatics (ISI), Berlin, Heidelberg*, May 2005, pp. 27–36.
- [11] S. Gregory, "An algorithm to find overlapping community structure in networks," in *Proceedings of Knowledge Discovery in Databases (PKDD), Berlin, Heidelberg*, September 2007, pp. 91–102.
- [12] —, "Finding overlapping communities in networks by label propagation," *New Journal of Physics*, vol. 12, no. 10, pp. 103 018–103 043, October 2010.
- [13] J. Xie, B. K. Szymanski, and X. Liu, "SLPA: Uncovering overlapping communities in social networks via a speaker-listener interaction dynamic process," in *Proceedings of the 11th IEEE International Conference on Data Mining (ICDM) Workshops, Vancouver, Canada*, September 2011, pp. 344–349.
- [14] M. Coscia, G. Rossetti, F. Giannotti, and D. Pedreschi, "DEMON: A local-first discovery method for overlapping communities," in *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Beijing, China*, August 2012, pp. 615–623.
- [15] J. Yang and J. Leskovec, "Overlapping community detection at scale: A nonnegative matrix factorization approach," in *Proceedings of the 6th ACM International Conference on Web Search and Data Mining (WSDM), Rome, Italy*, February 2013, pp. 587–596.
- [16] A. Holloco, T. Bonald, and M. Lelarge, "Multiple local community detection," *ACM SIGMETRICS Performance Evaluation Review*, vol. 45, pp. 76–83, March 2018.
- [17] Y. Li, K. He, K. Kloster, D. Bindel, and J. E. Hopcroft, "Local spectral clustering for overlapping community detection," *ACM Transactions on Knowledge Discovery from Data*, vol. 12, no. 2, pp. 17–(1–27), March 2018.
- [18] G. Rossetti, "Exorcising the demon: Angel, efficient node-centric community discovery," in *Proceedings of International Conference on Complex Networks and Their Applications VIII, Lisbon, Portugal*, November 2019, pp. 152–163.
- [19] D. Jin, Z. Jin, P. Jiao, S. Pan, D. He, J. Wu, P. Yu, and W. Zhang, "A survey of community detection approaches: From statistical modeling to deep learning," *IEEE Transactions on Knowledge and Data Engineering*, pp. 1–22, August 2021.
- [20] S. Fortunato and D. Hric, "Community detection in networks: A user guide," *Physics Reports*, vol. 659, pp. 1–44, 2016.
- [21] M. A. Javed, M. S. Younis, S. Latif, J. Qadir, and A. Baig, "Community detection in networks: A multidisciplinary review," *Journal of Network and Computer Applications*, vol. 108, pp. 87–111, 2018.
- [22] P. Bedi and C. Sharma, "Community detection in social networks," *WIREs Data Mining and Knowledge Discovery*, vol. 6, no. 3, pp. 115–135, 2016.
- [23] I. Derényi, G. Palla, and T. Vicsek, "Clique percolation in random networks," *Physical Review Letter*, vol. 94, pp. 160 202–(1–4), April 2005.
- [24] U. N. Raghavan, R. Albert, and S. Kumara, "Near linear time algorithm to detect community structures in large-scale networks," *Physical Review E*, vol. 76, no. 3, pp. 036 106–(1–11), October 2007.
- [25] Y. Gao, X. Yu, and H. Zhang, "Overlapping community detection by constrained personalized pagerank," *Expert Systems with Applications*, vol. 173, pp. 114 682–(1–12), February 2021.
- [26] O. Doluca and K. Oğuz, "APAL: Adjacency propagation algorithm for overlapping community detection in biological networks," *Information Sciences*, vol. 579, pp. 574–590, August 2021.
- [27] A. K. Ghoshal, N. Das, and S. Das, "Disjoint and overlapping community detection in small-world networks leveraging mean path length," *IEEE Transactions on Computational Social Systems*, vol. 9, no. 2, pp. 406–418, July 2021.
- [28] T. Ma, Q. Liu, J. Cao, Y. Tian, A. Al-Dhelaan, and M. Al-Rodhaan, "LGIEM: Global and local node influence based community detection," *Future Generation Computer Systems*, vol. 105, pp. 533–546, 2020.
- [29] J. Chen, L. Chen, Y. Chen, M. Zhao, S. Yu, Q. Xuan, and X. Yang, "GA-based Q-attack on community detection," *IEEE Transactions on Computational Social Systems*, vol. 6, no. 3, pp. 491–503, 2019.
- [30] J. Gulati and M. Abulaish, "A novel snowball-chain approach for detecting community structures in social graphs," in *Proceedings of the 2019 IEEE Symposium Series on Computational Intelligence (SSCI), Xiamen, China*, December 2019, pp. 2462–2469.
- [31] W. W. Zachary, "An information flow model for conflict and fission in small groups," *Journal of Anthropological Research*, vol. 33, no. 4, pp. 452–473, November 1976.
- [32] D. Lusseau, K. Schneider, O. J. Boisseau, P. Haase, E. Slooten, and S. M. Dawson, "The bottlenose dolphin community of doubtful sound features a large proportion of long-lasting associations - can geographic isolation explain this unique trait?" *Behavioral Ecology and Sociobiology*, vol. 54, pp. 396–405, January 2003.
- [33] M. Girvan and M. E. J. Newman, "Community structure in social and biological networks," *Proceedings of the National Academy of Sciences (PNAS)*, vol. 99, no. 12, pp. 7821–7826, January 2002.
- [34] M. E. J. Newman, "Finding community structure in networks using the eigenvectors of matrices," *Physical Review E*, vol. 74, no. 3, pp. 036 104–(1–19), October 2006.
- [35] P. M. Gleiser and L. Danon, "Community structure in jazz," *Advances in Complex Systems*, vol. 6, no. 04, pp. 565–573, December 2003.
- [36] R. Guimerà, L. Danon, A. Díaz-Guilera, F. Giralt, and A. Arenas, "Self-similar community structure in a network of human interactions," *Physical Review E*, vol. 68, no. 6, pp. 065 103–(1–4), January 2003.
- [37] D. J. Watts and S. H. Strogatz, "Collective dynamics of 'small-world' networks," *Nature*, vol. 393, no. 6684, pp. 440–442, July 1998.
- [38] S. Gregory, "An algorithm to find overlapping community structure in networks," in *Proceedings of European Conference on Principles of Data Mining and Knowledge Discovery (PKDD), Warsaw, Poland*, September 2007, pp. 91–102.
- [39] G. Palla, I. Derényi, I. Farkas, and T. Vicsek, "Uncovering the overlapping community structure of complex networks in nature and society," *Nature*, vol. 435, no. 7043, pp. 814–818, July 2005.
- [40] A. Lancichinetti, S. Fortunato, and F. Radicchi, "Benchmark graphs for testing community detection algorithms," *Physical Review E*, vol. 78, no. 4, pp. 046 110–(1–5), November 2008.

Improving the Computational Complexity of the COOL Screening Tool

Mohamed Ghalwash
IBM Research, NY, USA
Ain Shams University, Cairo, Egypt

Abstract—Autoimmune disorder, such as celiac disease and type 1 diabetes, is a condition in which the immune system attacks body tissues by mistake. This might be triggered by abnormality in the development of biomarkers such as autoantibodies, which are generated by unhealthy beta cells. Therefore, screening of such biomarkers is crucial for early diagnosis of autoimmune diseases. However, one of the fundamental questions of screening is when to screen subjects who might be at a higher risk of autoimmune disorder. This requires an exhaustive search to find the optimal ages of screening in retrospective cohorts. Very recently, a comprehensive tool was developed for screening in autoimmune disease. In this paper, we improved the computational time of the algorithm used in the screening tool. The new algorithm is more than 100 times faster than the original one. This improvement would help to increase the utility of the tool among clinicians and research scientists in the community.

Keywords—Software engineering; screening tool; autoimmune disorder

I. INTRODUCTION

Autoimmune disorder is a condition in which the immune system mistakenly attacks healthy body tissues in different organs of the body. For example, in type 1 diabetes, the immune system destroys the insulin-producing cells of the endocrine pancreas, which leads to insulin deficiency [1]. In celiac disease, eating gluten – a protein found in wheat, rye, and barley – causes the autoimmune system to damage the small intestine [2]. There are many factors involved in causing such diseases such as genes, environmental factors, drugs and/or chemicals. However, autoimmune disorder is often associated with a few circulating autoantibodies, which are abnormal antibodies generated by pathogenic β -cells, when targeting a tissue [3]. Autoantibodies are often precede the onset of the disease and, therefore, considered as a clinical biomarker of the autoimmune disorder. In type 1 diabetes and celiac diseases, there are four or five autoantibodies that are often used to assess the risk of developing the disease [4], [5].

Screening for autoantibodies – a group of serum tests to assess the presence of autoantibodies – is usually performed to detect the disease as early as possible so that a proper treatment or intervention can be administered. Therefore, frequent screening is of utmost importance to detect potential autoimmune disorder in subjects who in an apparently healthy population [6], [7]. Although frequent screening is beneficial for detecting subjects who are at a higher risk of the disease, it is cost inefficient and may also introduce harm for those who do not have the diseases by increasing the risk of overdiagnosis [8]. Therefore, one needs to find the *optimal*

ages for screening in order to balance between the benefit and the harm of multiple screening.

To find a proper screening schedule, one needs to do cross-sectional experiments on retrospective cohort to find the optimal ages for screening. The authors of a recent paper [9] proposed a tool, called the Collaborative Open Outcomes tool (COOL), that can be used to compute the quality performance of a given proposed screening schedule according to some measures that can be used to balance between the benefit and the harm of the screening schedule. However, computing these measures for a given schedule is a very time consuming task. In this paper, we propose to make these computations much faster. This proposed enhancement will increase the utility of the tool to compare multiple schedules to find the optimal (according to the given measures) screening schedule much faster.

II. METHOD

A. Data Structure

We explain the structure of the data used for defining the screening schedule. The data has biomarkers information for each subject. Each subject may visit the clinic multiple times and each time a blood sample is taken from the subject to assess the development of biomarkers. The value of each biomarker is either positive (the autoantibody is developed) or negative. It is worth mentioning that each subjects may have a different number of visits.

Notations: We use the upper case letter to define a matrix – a two dimensional array –, e.g. X , a boldface letter to define a vector, e.g. \mathbf{x} , and a italic letter to define an entry or element, e.g. x . $\mathbf{x}[i]$ represents the entry i of the vector \mathbf{x} . $X[i]$ represents the i^{th} row of the matrix X , and $X[i][j]$ represents the entry in the row i and column j .

Mathematically, let us define the data for a subject i as $\mathbf{x}_i = [(t_i^1, \mathbf{x}_i^1), (t_i^2, \mathbf{x}_i^2), \dots, (t_i^{T_i}, \mathbf{x}_i^{T_i})]$ where T_i is the number of visits for the subject i , t_i^j is the subject's age at the visit j , and $\mathbf{x}_i^j \in \{0, 1\}^M$ is the list of M biomarkers for the visit j . In addition, the information about whether and when the disease was developed is recorded. For simplicity, we assume that each subject either developed the disease within a predefined period of time from birth or the subject has been observed for the full period but has not developed the disease¹. If the subject has

¹The other case where the subject is partially observed and has not developed the disease (right censored subjects) can be handled using inverse probability censoring weights [10] but it is outside the scope of this paper.

developed the disease, the subject is not followed afterwards. y_i is the age when the disease was developed and -1 otherwise.

Example II.1. Let us assume that there are four subjects, 1, 2, 3, and 4. The data for these four subjects can be represented as

- $\mathbf{x}_1 = [(1, [0, 1, 0, 1]), (2.3, [1, 1, 0, 0]), (5.8, [0, 1, 0, 1]), (7.1, [1, 0, 1, 0])]$, $y_1 = 9$
- $\mathbf{x}_2 = [(2.4, [0, 0, 0, 1]), (6, [1, 0, 0, 1]), (9.2, [0, 0, 0, 1])]$, $y_2 = -1$
- $\mathbf{x}_3 = [(1.9, [0, 0, 0, 0]), (7.4, [0, 0, 1, 1])]$, $y_3 = 8$
- $\mathbf{x}_4 = [(0.6, [0, 1, 0, 0]), (4.7, [0, 0, 0, 1]), (6.4, [0, 0, 1, 1]), (10, [0, 0, 0, 0])]$, $y_4 = -1$

Subject 1 has a sequence of $T_1 = 4$ visits. Each visit has measurements for $M = 4$ biomarkers. The first visit was measured at age $t_1^1 = 1$ year and the second and the fourth biomarkers were positives while the other two biomarkers were negatives. The second visit was sampled at age $t_1^2 = 2.3$ years. We can see that the fourth biomarker turned to negative in the second visit while the first biomarker became positive. The subject has developed the disease at age $y_1 = 9$ years. The second subject has $T_2 = 3$ visits at ages 2.4, 6, and 9.2 years and has not developed the disease, i.e. $y_2 = -1$. We clearly see that each subject may have a different number of visits and these visits might be sampled at different ages. A graphical representation of these data is shown in Fig. 1.

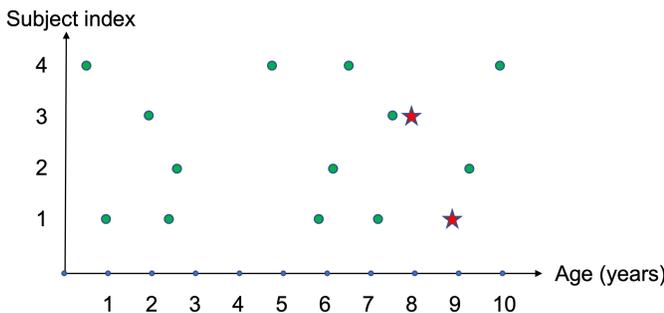


Fig. 1. A Graphical Representation for the given Data in the Example. The Green Points Represent Visits while the Red Stars Represent the Age at which the Subjects Developed the Disease. E.g., Subject 3 has 2 Visits and Developed the Disease at Age 8.

One simple data structure that can be used to store the data for all subjects is a 3-dimensional array, where the first dimension is the number of subjects N , the second dimension is the maximum number of visits $S = \max_i \{T_i\}$, and the third dimension is the number of biomarkers M , i.e. $\mathbb{R}^{N \times S \times M}$. However, there are two challenges to store the data in a three-dimensional array. The first challenge is that each subject may have a different number of visits. The second challenge is the irregularity in the biomarkers collections. As seen from the example, the biomarkers are collected at different and irregular time stamps. These two issues pose a challenge to store the data for all subjects in a 3-dimensional array, which assumes that the data are time-aligned. A better data structure for storing such information would be a 2-dimensional array (matrix) with a special structure.

Let us assume that $T = \sum_{i=1}^N T_i$ is the total number of

visits across all N subjects. We construct a matrix X with dimensions $T \times M + 2$, where each row represents one visit for a particular subject. The first column in the matrix represents the subject index, the second column is the age of the subject at the current visit, the other M columns are the values of the biomarkers. Data is sorted in ascending order by subject index and age. An additional array \mathbf{y} stores the age at which the subject developed the disease, i.e. $\mathbf{y}[i]$ is the age when the subject i developed the disease and -1 otherwise.

Example II.2. The matrix for the data in Example II.1 can be represented as

$$X = \begin{bmatrix} 1 & 1.0 & 0 & 1 & 0 & 1 \\ 1 & 2.3 & 1 & 1 & 0 & 0 \\ 1 & 5.8 & 0 & 1 & 0 & 1 \\ 1 & 7.1 & 1 & 0 & 1 & 0 \\ 2 & 2.4 & 0 & 0 & 0 & 1 \\ 2 & 6.0 & 1 & 0 & 0 & 1 \\ 2 & 9.2 & 0 & 0 & 0 & 1 \\ 3 & 1.9 & 0 & 0 & 0 & 0 \\ 3 & 7.4 & 0 & 0 & 1 & 1 \\ 4 & 0.6 & 0 & 1 & 0 & 0 \\ 4 & 4.7 & 0 & 0 & 0 & 1 \\ 4 & 6.4 & 0 & 0 & 1 & 1 \\ 4 & 10.0 & 0 & 0 & 0 & 0 \end{bmatrix} \in \mathbb{R}^{T \times M + 2},$$

$$\mathbf{y} = \begin{bmatrix} 9 \\ -1 \\ 8 \\ -1 \end{bmatrix}$$

As it can be seen, subject 1 has 4 rows in the matrix representing 4 visits, and subject 3 has two rows.

B. Single-Age Screening

Problem 1 (Single-age screening). At which age a , subjects with a positive test at that age will likely develop the disease within the observation period?

The objective of screening at a single age is to assess the likelihood that a subject has the disease. Let us assume that the screening test is whether any biomarker is positive. The question would be how likely subjects with any positive biomarker at a given age will develop the disease within the observation period (e.g. within 10 years from birth). In order to compute the quality performance of the screening at a single age, we need to compute the following Table I:

TABLE I. SUMMARY OF THE SCREENING TEST RESULTS

Screening test	Developed the disease	Not developed the disease
Positive	# true positives (TP)	# false positives (FP)
Negative	# false negatives (FN)	# true negatives (TN)
No test	# no test and positives (NP)	# no test and negatives (NN)

Each subject will be placed in one of these six cells. If the subject was tested positive and developed the disease, the subject will be counted in the TP cell. FP is the number of subjects who were tested positive and have not developed the disease. Similarly, FN (TN) is the number of subjects

who were tested negative and developed (not developed) the disease, respectively. Finally, since not all subjects may not necessarily have a visit at a particular age, some subjects may have no screening test and therefore will be missing from the screening test. This is accounted for in the last row of Table I.

Using the information provided in Table I, the screening test is usually evaluated using the sensitivity and the specificity measures [11]. The sensitivity is the probability that the screening test is positive among those who have the disease. Specificity is the probability that the screening test is negative among those who do not have the disease [12]. These two measures can be computed as:

$$Sen = \frac{TP}{TP + FN} \quad (1)$$

$$Spc = \frac{TN}{TN + FP} \quad (2)$$

Example II.3. If the sensitivity is 80%, it means that 80% of diseased subjects are identified as diseased (have a positive test). If the specificity is 90%, it means that 90% of non-diseased subjects have a negative test (correctly identified as non-diseased).

These two measures are important as they measure the percentage of diseased individuals who have positive test results and the percentage of non-diseased individuals who have negative test results, respectively. Nevertheless, these two measures assume that the test result for each subject is known, i.e. they do not account for subjects with missing tests. Cumulative sensitivity ($CSen$) and dynamic specificity ($DSpc$) address this issue [13]:

$$CSen = \frac{TP}{TP + FN + NP} \quad (3)$$

$$DSpc = \frac{TN}{TN + FP + NN} \quad (4)$$

As it can be seen, $CSen$ and $DSpc$ require all subjects who do/do not have the disease, respectively. However, from the subject's perspective, these two measures do not give insights about the likelihood to develop the disease if the test results is positive or negative. Positive predictive value (PPV) and negative predictive value (NPV) answer this question.

$$PPV = \frac{TP}{TP + FP} \quad (5)$$

$$NPV = \frac{TN}{TN + FN} \quad (6)$$

PPV is the probability of having the disease among those subjects who tested positive. NPV is the probability of not having the disease among those subjects who tested negative.

So, in order to evaluate the performance of a screening test, we need to compute 4 measures $CSen$, $DSpc$, PPV and NPV . Algorithm 1 evaluates the performance of a screening at a given age a by computing these four measures.

Algorithm 1 takes as parameters the age a at which the screening will be evaluated, the data matrix X that encodes the age and the biomarkers information, and the label array y that encodes the age at which the disease was developed. The algorithm utilizes an array $found$ to mark whether the subject

Algorithm 1: Single-Age Screening (SS)

Input: Age a , encoding data matrix X , label array y
Return: $CSen$, $DSpc$, PPV , and NPV .
// list for all N subjects
1 Initialize all N entries of the $found$ list with false
Initialize TP , TN , FP , FN , NP , and NN with zeros.
2 **for** each row in X **do**
 // row is a list of $M+2$ entries
3 $id = row[1]$
4 $age = row[2]$
5 $biomarkers = row[3 : M + 2]$
 /* if the age is within 6 months of a */
6 **if** $a - .5 \leq age < a + .5$ **then**
 // found a visit for the current subject
7 $found[id] = true$
8 **if** $IsPos(biomarkers)$ **then**
9 $PositiveTest(y[id])$
10 **else**
11 $NegativeTest(y[id])$
 /* loop over subjects with a missing screening test to compute NN and NP . */
12 **for** each subject id **do**
 // if the test is missing
13 **if** $found[id]$ is false **then**
14 $MissingTest(y[id])$
15 Compute $CSen$, $DSpc$, PPV , NPV using equations 3-6

Algorithm 2: Helper Functions

1 **Function** $PositiveTest(label)$:
2 **if** $label \geq 0$ **then**
3 $TP = TP + 1$ // diseased subject
4 **else**
5 $FP = FP + 1$ // non-diseased subject
6 **return**
7 **Function** $NegativeTest(label)$:
8 **if** $label \geq 0$ **then**
9 $FN = FN + 1$ // diseased subject
10 **else**
11 $TN = TN + 1$ // non-diseased subject
12 **return**
13 **Function** $MissingTest(label)$:
14 **if** $label \geq 0$ **then**
15 $NP = NP + 1$ // diseased subject
16 **else**
17 $NN = NN + 1$ // non-diseased subject
18 **return**
19 **Function** $IsPos(biomarkers)$:
20 **return**

has a screening test at the given age a , i.e. $found[i] = 1$ if the subject i has a visit at the given age a , and 0 otherwise.

In line 1, the algorithm initializes the boolean array *found* with false. In line 2, it initializes all counts with zero. Then, it loops over all rows in the data matrix *X* (line 3), and for each row it checks whether the age of the current visit is within a specified window of 6 months around the given age (line 7). If yes, it marks that the subject has been tested (line 8) and checks the results for the screening test (line 9) using the function *IsPos*. If the test result is positive (line 10), the algorithm calls the function *PositiveTest* in Algorithm 2, which updates the number of true positives or false positive depending on whether the patient has developed the disease. Otherwise, it updates the number of false positives (line 12). If the test result is negative (line 11), the algorithm calls the function *NegativeTest* which updates either false negatives if the patient developed the disease or true negatives if the patient has not developed the disease.

Finally, after iterating over the entire matrix *X*, the algorithm iterates over the *found* array (line 13) to find those who have not been tested at the given age (line 14) and calls the function *MissingTest* in line 15 to compute the number of subjects who missed the screening test and developed (*NP*) or did not develop the disease (*NN*). After computing *TP*, *TN*, *FP*, *FN*, *NP*, and *NN* counts, the algorithm uses equations (3-6) to compute *CSen*, *DSpc*, *PPV*, and *NPV* for the single-age screening at age *a*.

Time Complexity: The for loop in line 3 has $O(T)$ iterations. Let us assume that the function *IsPos* in line 9 takes $O(M)$. The loop in line 18 takes $O(N)$. Hence, the total time complexity of Algorithm 1 is $O(T.M + N)$.

Example II.4. We compute the quality performance of screening for any biomarker (if any biomarker is positive, the result of the test is positive) at age 2 using data provided in Example II.2. The summary statistics of screening at age 2 is given in the following Table II:

TABLE II. SUMMARY STATISTICS FOR SCREENING OF ANY BIOMARKER AT AGE 2. THE SUBJECT ID COLUMNS INDICATES THE SUBJECTS USED FOR COMPUTING THE MEASURE OR THE COUNT

Screening test	Count	Subject ids
<i>TP</i>	1	1
<i>FP</i>	1	2
<i>TN</i>	0	
<i>FN</i>	1	3
<i>NP</i>	0	
<i>NN</i>	1	4
<i>CSen</i>	0.5	1,3
<i>DSpc</i>	0	2,4
<i>PPV</i>	0.5	1,2
<i>NPV</i>	0	3

The screening at a single age might not perform good as some subjects might miss the screening test and that will reduce the sensitivity and/or specificity of the test. To increase the quality performance of a screening, one can screen twice so that those subjects who missed the first screening can be

covered by the second screening. This is discussed in the next section.

C. Two-Age Screening

Problem 2 (Two-age screening). How likely subjects with a positive test at either one of a pair of ages *a* and *b* will develop the disease within the observation period?

The screening test can be performed at the first age *a*. If the result is positive then no need to screen again and the final result is positive. If the screening test is negative or the subject missed the first screening then another screening is required at the second age and the result of the second screening determines the final result. If the subject missed both screening then it will be counted either in *NN* or *NP* depending whether the subject developed the disease. The two-age screening can be visualized as in Fig. 2.

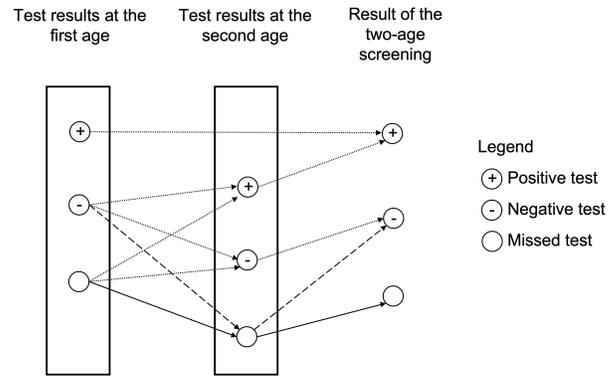


Fig. 2. Visualization of the Two-Age Screening Process. The Final Result of the Screening is Positive if and only if One of the Screenings at the First or the Second Age is Positive. The Final Result is Missing if both Screenings are Missing. Otherwise, the Final Result is Negative.

Algorithm 3 describes the two-age screening process. The algorithm takes a pair of ages *a* and *b* to compute the screening results where $a < b$. For each row in *X*, it tests whether the current visit are within the window of 6 months of age *a* (line 7). If the subject has a visit within that window, the algorithm applies the screening test (line 8) and if the result is positive, it marks that the subject *id* has a positive test at age *a* (line 9) and then updates *TP* and *TN* in line 10. If the result is negative, it marks that the subject has a negative first screening (line 12).

If the current visit is not within the window of 6 months around *a*, the algorithms checks for the second screening (note that the matrix *X* is sorted in ascending order by age). If the visit is within the window of 6 months around the second age *b* and if the subject has no positive results in the first screening (line 13), then it checks the results of the screening at the second age (line 14). If the screening at the second age is positive, the algorithm marks that the second screening is positive (line 15) and updates the counts *TP* and *FP* (line 16). If the second screening is negative, it marks that the second screening is negative (line 18).

After iterating over all rows in *X*, the algorithm iterates over all subjects who missed the first and the second tests (line

Algorithm 3: Two-Age Screening (TS)

Input: Ages a, b where $a < b$, encoding data matrix X , label array \mathbf{y}
Return: $CSen, DSpc, PPV$, and NPV .
// assume all subject missed both screenings

- 1 Initialize all N entries of $found1$ and $found2$ with -1
- 2 Initialize TP, TN, FP, FN, NP , and NN with 0
- 3 **for** each row in X **do**
- 4 $id = row[1]$
- 5 $age = row[2]$
- 6 $biomarkers = row[3 : M + 2]$
- 7 */* if the age is within the window of a */*
- 8 **if** $a - .5 \leq age < a + .5$ **then**
- 9 *// found a visit for the current subject*
- 10 **if** $IsPos(biomarkers)$ **then**
- 11 $found1[id] = 1$ *// 1st test is positive*
- 12 $PositiveTest(\mathbf{y}[id])$
- 13 **else**
- 14 $found1[id] = 0$ *// 1st test is negative*
- 15 **else if** $b - .5 \leq age < b + .5 \wedge found1[id] \neq 1$ **then**
- 16 **if** $IsPos(biomarkers)$ **then**
- 17 $found2[id] = 1$ *// test is pos*
- 18 $PositiveTest(\mathbf{y}[id])$
- 19 **else**
- 20 $found2[id] = 0$ *// test is neg*
- 21 */* loop over subjects with a missing test */*
- 22 **for** each subject id **do**
- 23 *// if both tests are missing*
- 24 **if** $found1[id] = -1 \wedge found2[id] = -1$ **then**
- 25 $MissingTest(\mathbf{y}[id])$
- 26 *// if 1st test is neg and 2nd is missing*
- 27 **else if** $found1[id] = 0 \wedge found2[id] = -1$ **then**
- 28 $NegativeTest(\mathbf{y}[id])$
- 29 Compute $CSen, DSpc, PPV, NPV$ using equations (3-6)

20) to update the counts NN and NP (line 21) and iterates over subjects who tested negative in the first screening and missed the second screening (line 22) to update the FN and TN counts (line 23). Finally, the screening quality measures are computed in line 24.

Time Complexity: The time complexity of Algorithm 3 is $O(TM + N)$.

Although the time complexity of Algorithm 3 is $O(TM + N) \approx O(T)$, but the actual running time is very large, especially if the algorithm needs to be executed multiple times. For example, in almost all cases in medical context, a confidence interval for each measure (sensitivity, specificity, PPV and NPV) is required. To compute the confidence interval [14], the algorithm needs to be run thousands of times on different samples of the matrix X . In addition, to compare different screening schedules, we compute the confidence interval for

each schedule and compare them to see how statistically significant the difference between the screening schedules is [15]. Therefore, it is preferred that the algorithm that computes the quality performance of the screening needs to be fast enough so that all these experiments can be run in a reasonable time.

To do that, we perform a data pre-processing that needs to be done only once, and then we will devise Algorithm 3 to make it faster which can be run multiple times and obtain the results much faster than using Algorithm 3.

D. Improved Two-Age Screening

We start with the improved algorithm for the two-age screening which can be easily modified for single-age screening. To improve the computational time of the two-age screening algorithm, we preprocess the data in a different data structure so that the computation becomes faster. The preprocessing step needs to be executed only once for the data and then each application of the two-age screening uses the preprocessed data and returns the results faster than the original algorithm.

For now, let us assume that we have already constructed a matrix B that contains the biomarker information, which will be used by the screening schedule algorithm (the construction of this matrix is explained in Section II-E). $B \in \mathbb{Z}^{N \times A}$ where A is the number of all possible distinct ages in the data that the screening are to be evaluated at, and N is the number of subjects. The entry $B[id][a] \in \{-1, 0, 1, 2, \dots, 2^M\}$ has the *encoding* of the biomarkers for the subject id at age a . Since the biomarkers are binary, then the number of all possible cases of biomarkers values is 2^M (note that the number of biomarkers is usually small in these applications as explained in the introduction section). The value -1 indicates that the subject id missed the test at age a .

Example II.5. Given Example II.2, there are $2^4 + 1 = 17$ possible values for each entry in the matrix B . The encoding matrix B is shown here:

$$B = \begin{bmatrix} 10 & 3 & -1 & -1 & -1 & 10 & 5 & -1 & -1 & -1 \\ -1 & 8 & -1 & -1 & -1 & 9 & -1 & -1 & 8 & -1 \\ -1 & 0 & -1 & -1 & -1 & -1 & 12 & -1 & -1 & -1 \\ 2 & -1 & -1 & -1 & 8 & 12 & -1 & -1 & -1 & 0 \end{bmatrix}$$

Column j encodes the biomarker information at age j . For example, the entry $B[2][6]$ encodes the biomarker information for subject $id = 2$ at age 6. The biomarker for subject 2 at age 6 were $[1, 0, 0, 1]$ which can be encoded as $2^1 + 2^0 + 2^0 + 2^1 = 9$. Similarly, the biomarker of subject 3 at age 2 is encoded as $B[3][2] = 2^0 + 2^0 + 2^0 + 2^0 = 0$. All entries with -1 indicate that the subject has no visit at that age, e.g. $B[2][5] = -1$ because the subject 2 has no visit at age 5.

Note that all ages are rounded given the window of interest. For example, visits at ages 2.4 are considered at age 2 (this is similar to line 6 in Algorithm 1)²

Given the matrix B , the improved algorithm for two-age screening is re-written in Algorithm 4. The algorithm iterates over all subjects (line 1), and for each subject id it checks if screening at age a and age b are missing (line 2) then it marks

²If there are multiple visits within the window around the given age, we can consider either the closest visit, the first visit, the last visit, or any other visit based on the application. This is outside the scope of this paper.

that the final result is missing (line 3). If one of the tests is positive (line 4) it marks that the final result is positive (line 5). Otherwise, it marks that the final results is negative (line 7).

Algorithm 4: Improved Two-Age Screening (ITS)

Input: Ages a, b where $a < b$, biomarkers matrix B , label array \mathbf{y}
Return: $CSen, DSpc, PPV$, and NPV .
 /* loop over all subjects */
 1 **for** each subject id **do**
 // if both tests are missing
 2 **if** $B[id][a] = -1 \wedge B[id][b] = -1$ **then**
 3 $MissingTest(\mathbf{y}[id])$
 // one of tests is positive
 4 **else if** $IsPos(B[id][a]) \vee IsPos(B[id][b])$ **then**
 5 $PositiveTest(\mathbf{y}[id])$
 // (both tests are negative) or (one is
 negative and the other is missing)
 6 **else**
 7 $NegativeTest(\mathbf{y}[id])$
 8 Compute $CSen, DSpc, PPV, NPV$ using equations (3-6)

Time Complexity: The running time for Algorithm 4 is $O(N)$.

E. Data Preprocessing for ITS

We preprocess the data only once to construct the biomarker encoding matrix B which makes the algorithm runs faster as evident by our experiments. The algorithm for constructing the matrix B is shown in Algorithm 5. The algorithm iterates over all rows of the matrix X (line 2). For each row, it maps the age to the closest age (line 6), encodes the biomarkers (line 7), and stores the value in the matrix B (line 8). To encode the biomarker information into one integer value (line 9), we multiple the biomarker vector into the encoding vector (line 10) to obtain the code value (line 11).

Algorithm 5: Biomarker Matrix

Input: matrix X
Return: biomarkers matrix B
 1 Initialize all entries of B with -1
 /* loop over all subjects */
 2 **for** each row in X **do**
 3 $id = row[1]$
 4 $age = row[2]$
 5 $biomarkers = row[3 : M + 2]$
 6 $a = Round(age)$ // map it to the closed age
 7 $code = encode(biomarkers)$
 8 $B[id][a] = code$
 9 **Function** $encode(biomarker)$:
 10 $e = [1 \ 2 \ 4 \ 16 \ \dots \ 2^M]^T$ // column
 vector
 11 $code = biomarker \times e$ // matrix
 multiplication
 12 **return** code

Time Complexity: The time complexity of Algorithm 5 is $O(T)$ but this process is executed only once not for each application of screening.

F. Improved Single-Age Screening

The improved algorithm for a single-age screening is shown in Algorithm 6.

Algorithm 6: Improved Single-Age Screening (ISS)

Input: Ages a , biomarkers matrix B , label array \mathbf{y}
Return: $CSen, DSpc, PPV$, and NPV .
 /* loop over all subjects */
 1 **for** each subject id **do**
 // if the test is missing
 2 **if** $B[id][a] = -1$ **then**
 3 $MissingTest(\mathbf{y}[id])$
 // the test is positive
 4 **else if** $IsPos(B[id][a])$ **then**
 5 $PositiveTest(\mathbf{y}[id])$
 // (the test is negative)
 6 **else**
 7 $NegativeTest(\mathbf{y}[id])$
 8 Compute $CSen, DSpc, PPV, NPV$ using equations (3-6)

III. EXPERIMENTS

We evaluated the performance of the SS, TS, ISS and ITS algorithms on datasets with different number of subjects and visits. The description of the datasets is shown in Table III. The experiments were run on a Mac laptop with processor 2.7 GHz Quad-Core Intel Core i7 and 16 GB of memory. The screening test used for these experiments is to test for any positive biomarker, i.e. if any biomarker is positive the result of the screening test is positive. The code is written in Python [16]. Python has a data structure called pandas dataframe [17] which can be used store information in the matrix X . Using the dataframe, the SS and TS algorithm can be even run faster if we filter the dataframe on rows where the age is within the 6 months window of the given age a . This is done using the command

$df[(df['age'] \geq a-0.5) \ \& \ (df['age'] < a+0.5)]$

In all our experiments for the SS and the TS algorithms we used the above command.

TABLE III. EACH SUBJECT HAS ON AVERAGE 30 VISITS. THERE ARE 3 BINARY BIOMARKERS. SUCH AS DISTRIBUTION OF VISITS, NUMBER OF SUBJECTS, ETC

# subjects	# total visits	average # visits
9.170	169,530	19
13.383	219,276	16
15.747	240,917	15
18.984	262,233	14

A. Single Age Screening

We compared the running time for the single-age screening algorithms SS and ISS on different datasets. The experiments were run multiple times and the median and quartiles of the running times are reported as shown in Fig. 3.

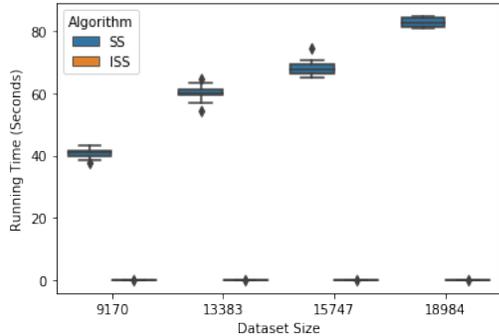


Fig. 3. Running Time Comparison between SS and ISS.

It is clear that the running time of the SS algorithm increases linearly with the dataset size. It takes about 80 seconds for Algorithm 1 to compute the quality performance of screening at a single age on data that has about 19,000 subjects, while the improved algorithm ISS takes only a fraction of a second to get the results. The running time for the ISS algorithm is shown in Fig. 4.

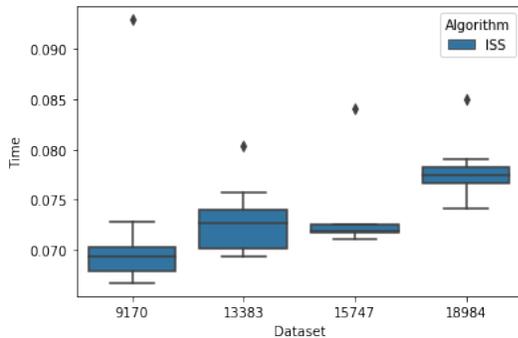


Fig. 4. Running Time for the ISS Algorithm.

B. Two Ages Screening

We compared the running time for the TS and ITS algorithms to compute the performance of the two-age screening. The results are shown in Fig. 5. A very similar behavior is observed. The TS algorithm scales linearly with the dataset size. The ITS algorithm is much faster than the TS algorithm. The running time for the ITS algorithm is shown in Fig. 6. ITS takes only 0.1 seconds to compute the quality performance of screening at a given two ages while TS takes 175 seconds.

C. Data Preprocessing for ISS and ITS

The additional overhead that the improved algorithms add on top of the original algorithms is the data preprocessing, i.e. the construction of the biomarkers encoding matrix B . This

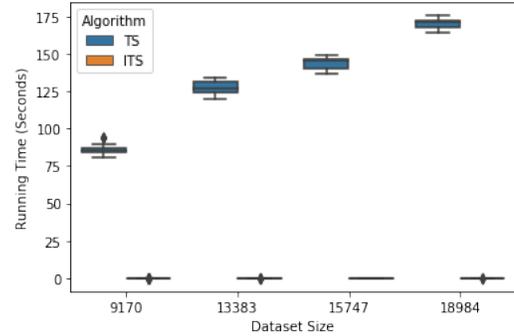


Fig. 5. Running Time Comparison between TS and ITS.

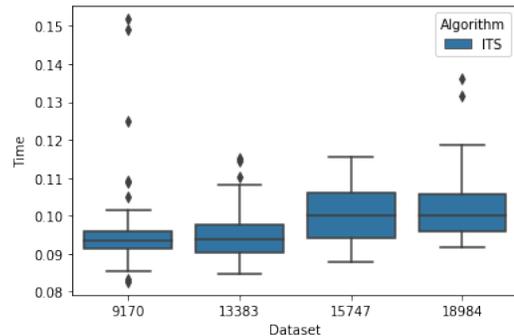


Fig. 6. Running Time for the ITS Algorithm.

step is required only once for each dataset. The running time for Algorithm 5 is shown in Table IV.

TABLE IV. RUNNING TIME FOR CONSTRUCTING THE MATRIX B

Dataset size (# subjects)	Time (mins)
9170	5.3
13383	6.7
15747	7.5
18984	8.4

IV. CONCLUSION

Screening of biomarkers is of utmost importance to assess the risk of developing autoimmune diseases such as type 1 diabetes and celiac diseases. To improve the quality performance of the screening test, screening more than one time is required. Algorithms to compute the quality performance of a screening schedule were developed as part of a screening tool. However, the running time of these algorithms are large which hinders the utility of the tool on large applications. We improved the running time of the screening algorithms by more than 800 times at an additional cost of preprocessing the data only once. We evaluated the running time of these screening algorithms on datasets with different sizes.

REFERENCES

- [1] S. A. Paschou, N. Papadopoulou-Marketou, G. P. Chrousos, and C. Kanaka-Gantenbein, "On type 1 diabetes mellitus pathogenesis," *Endocrine connections*, vol. 7, no. 1, pp. R38–R46, 2018.

- [2] P. H. Green and C. Cellier, "Celiac disease," *New england journal of medicine*, vol. 357, no. 17, pp. 1731–1743, 2007.
- [3] Z. X. Xiao, J. S. Miller, and S. G. Zheng, "An updated advance of autoantibodies in autoimmune diseases," *Autoimmunity Reviews*, vol. 20, no. 2, p. 102743, 2021.
- [4] C. E. Taplin and J. M. Barker, "Autoantibodies in type 1 diabetes," *Autoimmunity*, vol. 41, no. 1, pp. 11–18, 2008.
- [5] S. Caja, M. Mäki, K. Kaukinen, and K. Lindfors, "Antibodies in celiac disease: implications beyond diagnostics," *Cellular & molecular immunology*, vol. 8, no. 2, pp. 103–109, 2011.
- [6] W. H. Organization *et al.*, "Screening programmes: a short guide. increase effectiveness, maximize benefits and minimize harm," 2020.
- [7] L. Frommer and G. J. Kahaly, "Type 1 diabetes and associated autoimmune diseases," *World journal of diabetes*, vol. 11, no. 11, p. 527, 2020.
- [8] N. Gilbert, "The pros and cons of screening," *Nature*, vol. 579, no. 7800, pp. S2–S2, 2020.
- [9] M. Ghalwash, E. Koski, R. Veijola, J. Toppari, W. Hagopian, M. Rewers, and V. Anand, "Simulating screening for risk of childhood diabetes: The collaborative open outcomes tool (cool)," in *AMIA Annual Symposium Proceedings*, vol. 2021. American Medical Informatics Association, 2021, p. 516.
- [10] D. M. Vock, J. Wolfson, S. Bandyopadhyay, G. Adomavicius, P. E. Johnson, G. Vazquez-Benitez, and P. J. O'Connor, "Adapting machine learning techniques to censored time-to-event health record data: A general-purpose approach using inverse probability of censoring weighting," *Journal of biomedical informatics*, vol. 61, pp. 119–131, 2016.
- [11] S. Nissen-Meyer, "Evaluation of screening tests in medical diagnosis," *Biometrics*, pp. 730–755, 1964.
- [12] R. Trevethan, "Sensitivity, specificity, and predictive values: foundations, pliabilities, and pitfalls in research and practice," *Frontiers in public health*, vol. 5, p. 307, 2017.
- [13] A. N. Kamarudin, T. Cox, and R. Kolamunnage-Dona, "Time-dependent roc curve analysis in medical research: current methods and applications," *BMC medical research methodology*, vol. 17, no. 1, pp. 1–19, 2017.
- [14] B. Efron, *The jackknife, the bootstrap and other resampling plans*. SIAM, 1982.
- [15] P. Armitage, G. Berry, and J. N. S. Matthews, *Statistical methods in medical research*. John Wiley & Sons, 2008.
- [16] G. Van Rossum and F. L. Drake Jr, *Python reference manual*. Centrum voor Wiskunde en Informatica Amsterdam, 1995.
- [17] T. pandas development team, "pandas-dev/pandas: Pandas," Feb. 2020. [Online]. Available: <https://doi.org/10.5281/zenodo.3509134>

A Lightweight Verifiable Secret Sharing in Internet of Things

Likang Lu¹

College of Computer Science; College of Software
Inner Mongolia University
Hohhot, 010021, China

Jianzhu Lu^{2*}

Department of Computer Science
Jinan University
Guangzhou, 510630, China

Abstract—Verifiable Secret Sharing (VSS) is a fundamental tool of cryptography and distributed computing in Internet of Things. Since network bandwidth is a scarce resource, minimizing the number of verification data will improve the performance of VSS. Existing VSS schemes, however, face limitations in meeting the number of verification data and energy consumptions for low-end devices, which make their adoption challenging in resource-limited IoTs. To address above limitations, we propose a VSS scheme according to Nyberg’s one-way Accumulator for one-way Hash Functions (NAHFs). The proposed VSS has two distinguished features: first, the security of the scheme is based on NAHFs whose computational requirements are the basic criteria for known IoT devices and, second, upon receiving only one verification data, participants can verify the correctness of both their shares and the secret without any communication. Experimental results show that, compared to the Feldman scheme and Rajabi-Eslami scheme, the energy consumption of a participant in the proposed scheme is respectively reduced by at least 24% and 83% for a secret.

Keywords—Verifiable secret sharing; one-way function; internet of things; security

I. INTRODUCTION

The Internet of Things (IoT) is moving at such a rapid pace that there is rising demand for transforming our physical world into a complex and dynamic system of connected devices. These IoT devices will be widely used in smart homes, body/health monitoring, environmental monitoring, condition-based maintenance, among many others. IoT is not a single technology. It is a combination of sensors, devices, networks, and software that works on a collaborative basis to achieve a common goal. Secure and reliable group communication has become critical in the IoT system. Group key agreement is widely employed for secure group communications in modern collaborative and group-oriented applications. The central challenge is secure and efficient group key management [1, 2]. This is because these IoT devices have limited computing ability and the limitation of communication bandwidth. In this paper, we focus on the design of lightweight verifiable secret sharing (VSS) schemes in order to achieve the secret reconstruction among a set of IoT devices, where the reconstructed secret may be the group key of them.

A. Motivation for Lightweight VSS

To date, there are two main families of approaches that have been investigated to provide VSS to participants. The

first approach provided verification data based on public key cryptography such as ASPP [3] in cyclic lattices and DLP [4]. The second approach to add verification capabilities to a scheme, was to use one-way functions to obtain fingerprints/signatures of the involved data [5]. However, the existing schemes suffer from some major problems. Firstly, existing schemes face the challenge in very large-scale deployment of IoT devices. Since verification data grew linearly with either the number of participants [5] or the threshold value [3], their performance dropped sharply as the number of IoT devices grows. Note that network bandwidth is a scarce resource. Minimizing the number of public verification data will improve the performance of VSS. In this paper, we address this challenge and propose a VSS Scheme with only one verification data used to verify a secret and all of its shares.

In addition, for these low-cost, battery-powered IoT devices, the lightweight implementation of VSS schemes has emerged as a critical issue. Because public key cryptography uses some big integers to generate the verification data, it is much slower than symmetric key cryptography, requires more processing power, and generally increases energy consumptions of participants [6]. When the batteries are low, it may cause the IoT devices to function abnormally. Existing solutions require the public-key computation (e.g., Modular exponentiation) that is an expensive operation for IoT devices in real systems. In the VSS setting, it is a challenge to design a lightweight VSS scheme that minimizes the energy consumption of a participant. To our knowledge, this paper represents the first effort in this direction.

B. Our Contribution

In this paper, we propose a lightweight VSS scheme in IoT environments. The security of the proposed scheme is based on NAHFs which are implemented through the generic symmetry-based hash function and simple bit-wise operation. The proposed scheme dictates to generate only a NAHF value as the verification data which proves the validity of the shares for all participants. Thus, the communication cost of each participant is reduced. In addition, each participant validates a received share by running an NAHF operation. Hence, the proposed scheme is computationally efficient for each participant. Furthermore, the computation and communication costs of each participant remain unchanged when the number of participants increases. That is, the proposed scheme provides the good scalability. Compared to the Feldman scheme [4] and Rajabi-Eslami scheme [3], the energy consumption of a

* Corresponding Author.

participant in the proposed scheme is respectively reduced by at least 24% and 83% for a secret. To the best of our knowledge, the approach of this paper is the first such technique that the number of verification data is only one value in the VSS scheme.

The rest of the article is structured as follows. Related work is presented in Section II, Section III presents a brief review of NAHF, Shamir's (t, n) secret sharing and VSS. Section IV is dedicated to the proposed VSS scheme including the security model, construction and security aspects. The performance analysis and simulation experiments for the proposed scheme are respectively discussed in Section V and Section VI. Section VII concludes the paper.

II. RELATED WORK

The secret sharing (SS) scheme is used as a tool in IoT applications including continuous authentication [1] and key management in sensor networks [7]. Such a scheme allows one to share a secret s among a set P of participants. The participants are assigned different values called shares and only certain authorized subsets of them were able to recover the secret using these shares. A (t, n) threshold SS scheme was introduced by Shamir [8] and Blakley [9] independently in 1979. In such a scheme, the authorized subsets consisted of all subsets of P including at least t participants. The scheme was unconditionally secure which meant that less than t participants found no information about the secret even with unlimited time and computing power. Then, many versions of SS were proposed to add some new features in the literatures [10].

A verifiable secret sharing (VSS) scheme is a generalization of a SS scheme [11], whose novelty is that everyone can verify whether the received share is a valid piece of the secret or not. The concept of VSS was first introduced by Chor et al [12] in 1985. Subsequently, based on "k-consistent" shares and interactive proof in [13], a VSS scheme was proposed to check the honesty of participants at the secret reconstruction phase. However, at the share generation phase, participants were unable to verify whether the shares they received from the dealer were valid. In 1987, a practical non-interactive VSS was proposed by Feldman [4, 5] through a homomorphic one-way function v for verifying consistency of each share. Indeed, let v be a $(+, \cdot)$ -homomorphic one-way function (that is, $v(a + b) = v(a) \cdot v(b)$); then, if v was evaluated over a polynomial $f(x) = \sum_{i=0}^{t-1} a_i x^i$, the equation $v(f(x)) = \prod_{i=0}^{t-1} v(a_i x^i)$ held. The dealer chose two primes p, q as public values and a generator g of a subgroup of order q of \mathbb{Z}_p^* , where q divided $p - 1$, and q was the lowest possible integer satisfying $g^q \equiv 1 \pmod{p}$. Then, it generated a share $s_j = f(x_j) \pmod{q}$ for each participant P_j , and published the public verification coefficients $A_i = g^{a_i} \pmod{p}$. Hence, the consistency of a share s_j was verified by checking the equality $g^{s_j} = \prod_{i=0}^{t-1} A_i^{x_j^i} \pmod{p}$. Here, the homomorphic property of exponentiation function $v(a) = g^a \pmod{p}$ was used. In the case of Feldman's scheme, the security was based on the hardness of the discrete logarithm problem (DLP). In 2019, Rajabi and Eslami [3] proposed a generic threshold VSS construction, and then presented a non-interactive VSS with security based on hardness of the approximate shortest polynomial problem

(ASPP) in cyclic lattices. In the work of Tsaloli et al. [14], by combining three different primitives (i.e., homomorphic hash functions, linearly homomorphic signatures, and threshold RSA signatures) as the baseline, an approach was proposed for protecting the secret data of clients and achieving public verifiability of the computed result. Recently, Koikara et al. [15] used a bilinear map to propose a publicly verifiable secret sharing (PVSS) scheme based on 3D-cellular automata. The VSS with bilinear pairings is not suitable for IoT systems because bilinear pairings are not friendly to lightweight devices [16]. In addition, the symmetry-based VSS is more suitable for the ultra-low energy devices as compared with the public key cryptographic approaches.

A new non-trapdoor accumulator for cumulative hashing was introduced by Nyberg [17]. This kind of accumulator is called a Nyberg's one-way Accumulator for one-way Hash Function (NAHF). In practice, the NAHF is effectively implemented by using the generic symmetry-based hash function and simple bit-wise operations. Oftentimes, this results in less memory requirements than digital signature-based solutions for verification problems. In 2017, Huang et al. [18] proposed a lightweight authentication scheme with dynamic group members in IoT environments. Here, based on a public secure NAHF, the proxy computed two accumulated hash values, W and R , which were used to verify whether the node was available and unrevoked. Recently, Fan et al. [19] presented a secure region-based handover scheme with user anonymity and fast revocation, where the region secret keys of the revoked users were accumulated by NAHFs. In the proposed scheme, the dealer generates the verification data with a NAHF such that the shares of participants can be publicly and efficiently verified. This enables us to add verification capability for participants using only one verification data.

III. PRELIMINARIES

In this section, we introduce some basic concepts of hash function, NAHF, secret sharing and VSS needed later

A. Notations

We shall use the following notations throughout the paper. A set with integers $1, 2, \dots, n$, is written either $\{1, 2, \dots, n\}$ or simply $[n]$. We denote by $|x|$ the length of the binary string corresponding to x , and $\lceil x \rceil$ the least integer that is greater than or equal to the given number x . Let $P = \{P_1, P_2, \dots, P_n\}$ be a set of n participants and D be the dealer. The threshold is denoted by t . Let $\mathbb{Z}_p, \mathbb{Z}_q$ be two finite fields and $\mathbb{Z}_q^* = \mathbb{Z}_q \setminus \{0\}$, where p is a prime modulus, q is a prime divisor of $p - 1$, and $q \geq n + 1$. We let $H : \{0, 1\}^r \times \{0, 1\}^* \rightarrow \{0, 1\}^r$ denote a Nyberg accumulated hash function, $h : \{0, 1\}^* \rightarrow \{0, 1, \dots, q - 1\}$ and $\hat{h} : \{0, 1\}^* \rightarrow \{0, 1\}^{rd}$ be two one-way hash functions, where h is used to construct the required H , and $r = |q|$.

B. Nyberg's One-way Accumulator for One-way Hash Function

In this paper, we review the concept of Nyberg's one-way Accumulator for one-way Hash Function (NAHF).

Definition 1 (One-way hash function [17]). *A family of one-way hash functions is an infinite set of functions $h_l : K_l \times S_l \rightarrow V_l$ having the following properties:*

- (1) There exists a polynomial P' such that for each integer l , $h_l(k, s)$ is computable in time $P'(l, |k|, |s|)$ for all $k \in K_l$ and all $s \in S_l$.
- (2) There is no polynomial P' such that there exists a probabilistic polynomial time algorithm which, for all sufficiently large l , when given l , a pair $(k, s) \in K_l \times S_l$, and a $s' \in S_l$, find an $k' \in K_l$ such that $h_l(k, s) = h_l(k', s')$ with probability greater than $1/P'(l)$, where (k, s) is chosen uniformly among all elements of $K_l \times S_l$ and s' is chosen uniformly from S_l .

Definition 2 (Quasi-commutativity [17]). A function $h : K \times S \rightarrow X$ is said to be quasi-commutative if for all $k \in K$ and for all $s_1, s_2 \in S$, $h(h(k, s_1), s_2) = h(h(k, s_2), s_1)$.

Definition 3 (Nyberg's one-way accumulator [17]). A family of one-way accumulators is a family of one-way hash functions with quasi-commutativity. The one-way accumulator by Nyberg [17] is constructed based on the generic symmetry-based hash function (e.g., SHA) and simple bit-wise operations. Compared to Benaloh's scheme [20], Nyberg's scheme is more efficient without employing asymmetric cryptographic operations.

Assume that $N = 2^d$ is an upper bound to the number of items to be accumulated and r is an integer. Let s_1, s_2, \dots, s_n be the accumulated items with different string sizes, and a set of the accumulated items $S = \{s_1, s_2, \dots, s_n\}$, where $n \leq N$. Assume that $H(\cdot, \cdot)$ denotes an NAHF from $\{0, 1\}^r \times \{0, 1\}^*$ to $\{0, 1\}^r$, and \odot is the bitwise operation AND. The NAHF is based on the one-way hash function $h : \{0, 1\}^* \rightarrow \{0, 1\}^{rd}$. All that is required to specify the NAHF is hashing process and AND operation. The heart of an NAHF is the hashing process. The hashing process applies a hash function h to the input to produce a r -bit output. The hashing process is composed of the following operations.

- Hashing operation: Hash the accumulated item s_i of the input and output a rd bits binary string $v_i = h(s_i)$.
- Transfer α : the NAHF does a transfer operation on the binary string v_i which is divided into r blocks, $(v_{i,1}, \dots, v_{i,r})$, of length d . The transfer of a block from a d -bit input to a bit output is performed as follows: If $v_{i,j}$ is a string of zero bits, it is replaced by 0; otherwise, $v_{i,j}$ is replaced by 1. That is, $\alpha(v_i) = (b_{i,1}, \dots, b_{i,r})$, where $b_{i,j} \in \{0, 1\}$, $j=1, \dots, r$.

In this way, we can transfer the accumulated item s_i to a bit string, $b_i = \alpha(h(s_i)) \in \{0, 1\}^r$, which can be considered as the values of r independent binary random variables if h is an ideal hash function.

The NAHF on an accumulated item $s_i \in S$ with an accumulated key $k \in \{0, 1\}^r$ can be implemented using the AND operation described as $H(k, s_i) = k \odot \alpha(v_i) = k \odot \alpha(h(s_i))$. And it also can be represented as $Z = H(k, s_i) = k \odot \alpha(v_i) = k \odot \alpha(h(s_i))$ for an accumulated item $s_i \in S$ ($i \in [n]$). The proposed VSS relies on the following properties of the NAHF $H(\cdot, \cdot)$:

- Quasi-commutativity: $H(H(k, s_1), s_2) = H(H(k, s_2), s_1)$.
- Absorbency: $H(H(k, s_i), s_i) = k \odot \alpha(h(s_i)) = H(k, s_i)$.

- An item s_i within the accumulated value Z can be verified by $H(Z, s_i) = Z \odot \alpha(h(s_i)) = Z$.

C. Shamir's Threshold Secret Sharing

There are n participants, $P = \{P_1, P_2, \dots, P_n\}$ and a dealer D . In Shamir's secret sharing scheme [8], it consisted of two phases: the share distribution phase and the secret reconstruction phase. During share distribution, the secret was $s = f(0)$, where $f(x)$ was a polynomial of degree $t - 1$ with random coefficients (except for the constant term), computed over a finite field. The participant $P_j \in P$ holding shares knew $s_j = f(x_j)$, where x_j was P_j 's unique nonzero identifier, $j \in [n]$. In secret reconstruction, any t out of n participants, P_{j_1}, \dots, P_{j_t} , were able to recover the secret s by using the Lagrange interpolation formula (1) or solving the following linear equations (2), where

$$s = f(0) = \sum_{i=1}^t s_{j_i} \left(\prod_{r=1, r \neq i}^t \frac{-x_{j_r}}{x_{j_i} - x_{j_r}} \right), \quad (1)$$

and

$$\begin{aligned} s_{j_1} &= s + a_1 \times x_{j_1} + \dots + a_{t-1} \times x_{j_1}^{t-1}, \\ s_{j_2} &= s + a_1 \times x_{j_2} + \dots + a_{t-1} \times x_{j_2}^{t-1}, \\ &\vdots \\ s_{j_t} &= s + a_1 \times x_{j_t} + \dots + a_{t-1} \times x_{j_t}^{t-1}. \end{aligned} \quad (2)$$

Note that the above coefficient matrix is a square Vandermonde matrix, which is invertible, since the x_j s are distinct.

D. VSS

In a SS scheme, participants must trust that shares they receive are correct. In a VSS scheme, additional verification data are given that allow each participant to check whether its share is correct. Each message that must be checked contains additional verification data. The verification data are sent in the clear, and can be used by the recipient to determine whether the share in the message is correct. That is, recipients use them to check that a point, (x_j, s_j) , sent to it is on the polynomial $f(x)$ and that the polynomial, $f(x)$, used as the basis for the sent shares equals the secret at $x=0$. The VSS is able to resist the following two kinds of active attacks: (1) some shares are tampered before being sent to the participants in the secret distribution phase; (2) participants submit error shares to others in the secret reconstruction phase.

IV. A LIGHTWEIGHT (t, n) VSS SCHEME

In the section, a lightweight (t, n) VSS scheme is proposed. We discuss techniques involving the security model, construction and the security aspects of the proposed scheme.

A. The Security Model of Proposed Scheme

In this section, we give the definition of a noninteractive (t, n) VSS scheme. There are n participants, $P = \{P_1, P_2, \dots, P_n\}$, and a dealer D . In the definition, there are four algorithms: share generation(SG), share verification(SHV), secret reconstruction(SR) and secret verification

(SEV). The proposed scheme consists of the share distribution phase and the secret reconstruction phase. We define a noninteractive (t, n) VSS scheme as follows:

A noninteractive (t, n) VSS scheme is a pair (share generation, secret reconstruction) of phases as follows.

- Share distribution: In this phase, on input a secret s and P_j 's identity x_j , D first runs SG algorithm to output a share for each participant and some verification data, where the shares are sent to the corresponding participants through a secure channel. Then, on input verification data and his share, each participant runs SHV algorithm to output accept or reject the share.
- Secret reconstruction: The input of this phase are the shares corresponding to a subset of participants. At first, the validity of each share is verified by other cooperating participants running SHV algorithm. Then, if the number of participants with valid shares is at least t , the secret can be computed by applying SR algorithm on the provided shares, and the recovered secret is verified by running SEV algorithm.

A non-interactive (t, n) VSS Scheme is called secure if it satisfies the following properties:

- Threshold. Every secret can only be recovered by any t or more participants who have received the shares, and any subset of participants with less than t participants cannot obtain any information about the secret.
- Verifiability/reconstructability: Every participant can verify his share in the share generation phase. During the secret reconstruction phase, the participants can validate the received shares and check if a reconstructed secret is correct.
- Security. The VSS scheme must be able to resist up to $t - 1$ colluded inside adversaries. In addition, any outside adversary cannot impersonate to be a member by forging a valid value after knowing at most $t - 1$ values from other members. The VSS scheme is secure, if the adversary cannot obtain the shares in polynomial time.

In addition, the following properties for a VSS are very much tailored to IoT devices as participants:

- Efficiency. The proposed scheme should have low calculation requirements and low communication costs at the participants to reduce their energy consumptions. This makes VSS for implementation on battery-powered IoT devices that have limited computing power.
- Scalability. Even if the number of participants in large-scale deployments is big, the communication cost of the scheme should be kept small to reduce the cost of the supporting network infrastructure.

B. The Proposed (t, n) VSS Scheme

Figure 1 shows the proposed (t, n) VSS scheme, where the combiner may be each participant in P . In the proposed

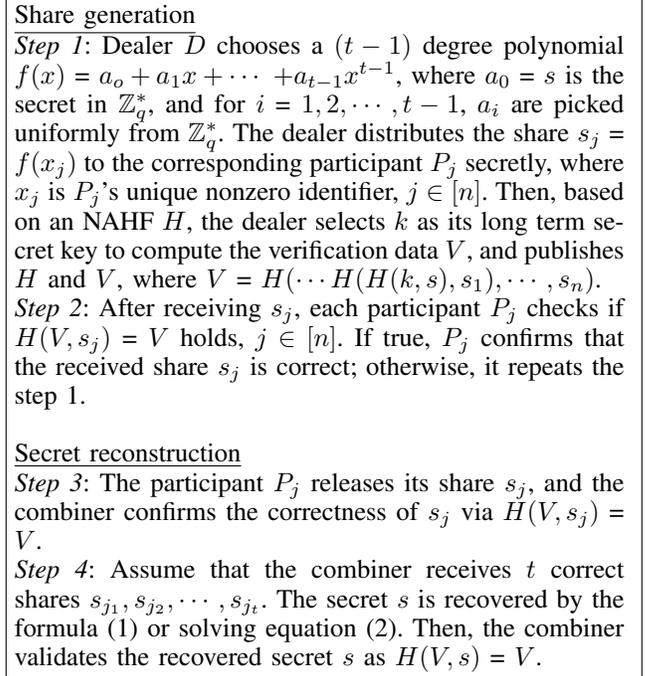


Fig. 1. The Proposed (t, n) VSS Scheme.

scheme, the algorithms SG, SHV, SR and SEV are the mathematical processes in the Step 1, 2, 3 and 4, respectively. The security of the scheme is based on an NAHF, which is quasi-commutative and has the absorbency property.

The correctness of the proposed (t, n) VSS scheme is guaranteed by the following theorem 1 and 2.

Theorem 1. In the share generation, the correctness of each share s_j can be validated by the receiver through $H(V, s_j) = V$, $j \in [n]$.

Proof 1. If the dealer D follows the scheme accurately, we have that $V = H(\dots H(H(k, s), s_1), \dots, s_n)$. Based on the absorbency property of H , it is known that the share s_n satisfies $H(V, s_n) = V$. In fact, $H(V, s_n) = H(H(\dots H(H(k, s), s_1), \dots, s_n), s_n) = H(\dots H(H(k, s), s_1), \dots, s_n) = V$, where the second equality holds for the absorbency property of H .

Generally, in accordance with the quasi-commutativity of H , we have

$$\begin{aligned}
 V &= H(\dots H(H(\dots H(H(k, s), s_1), \dots, s_j), s_{j+1}), \dots, s_n) \\
 &= H(\dots H(H(\dots H(H(k, s), s_1), \dots, s_{j+1}), s_j), \dots, s_n) \\
 &\quad \vdots \\
 &= H(H(\dots H(\dots H(H(k, s), s_1), \dots, s_{j+1}), \dots, s_n), s_j).
 \end{aligned} \tag{3}$$

where $j = 1, 2, \dots, n - 1$. Combining the absorbency property of H and equation (3), we obtain that $H(V, s_j) = H(H(H(\dots H(\dots H(H(k, s), s_1), \dots, s_{j+1}), \dots, s_n), s_j), s_j) = H(H(\dots H(\dots H(H(k, s), s_1), \dots, s_{j+1}), \dots, s_n), s_j) = V$, where the second equality holds for the absorbency property of H , and the third equality holds due to equation (3). This completes the proof.

Theorem 2. In the secret reconstruction, the received shares

s_{j_θ} and the recovered secret s can be publicly and efficiently verified via $H(V, s_{j_\theta}) = V$ and $H(V, s) = V$, respectively, $\theta \in [t]$.

Proof 2. In the secret reconstruction, the share s_{j_θ} can be publicly and efficiently verified via $H(V, s_{j_\theta}) = V$, for $\theta \in [t]$. This proof is the same as that of Theorem 1. In addition, similar to the derivation of equation (3), the secret s satisfies the following equation:

$$\begin{aligned} V &= H(\cdots H(H(k, s), s_1), \cdots, s_n) \\ &= H(\cdots H(H(k, s_1), s), \cdots, s_n) \\ &\vdots \\ &= H(H(\cdots H(H(k, s_1), s_2), \cdots, s_n), s). \end{aligned} \quad (4)$$

By using the absorbency property of H and equation (4), for the secret s we see that $H(V, s) = V$. This is because $H(V, s) = H(H(H(\cdots H(H(k, s_1), s_2), \cdots, s_n), s), s) = H(H(\cdots H(H(k, s_1), s_2), \cdots, s_n), s) = V$, where the second equality holds due to the absorbency property of H , and the third equality holds by equation (4). This completes the proof.

Remark 1. The correctness of algorithms $H(V, s_j) = V$ and $H(V, s) = V$ depends on the assumption that the output length, rd , of h satisfies $(n + 1) \leq 2^d$, where an NAHF $H : \{0, 1\}^r \times \{0, 1\}^* \rightarrow \{0, 1\}^r$ is constructed through $h : \{0, 1\}^* \rightarrow \{0, 1\}^{rd}$. When $(n + 1) > 2^d$, it is feasible to replace V with $(V_0, V_1, \cdots, V_{u-1})$, where $u = \lceil \frac{n+1}{2^d} \rceil$. For $\varsigma = 0, 1, \cdots, u-1$, $V^{(\varsigma)}$ is generated as follows: (1) different hash functions, $h^{(\varsigma)} : \{0, 1\}^* \rightarrow \{0, 1\}^{rd}$, are chosen. (2) the NAHF $H^{(\varsigma)} : \{0, 1\}^r \times \{0, 1\}^* \rightarrow \{0, 1\}^r$ is generated by the hash function $h^{(\varsigma)}$. (3) Let $s_{n+1} = s$, the ς -th value is computed as $V^{(\varsigma)} = H^{(\varsigma)}(\cdots H^{(\varsigma)}(k, s_{\varsigma+1}), \cdots, s_{\varsigma+2^d})$. To verify the correctness of $s_{\varsigma+j}$, we can check if $H^{(\varsigma)}(V^{(\varsigma)}, s_{\varsigma+j}) = V^{(\varsigma)}$, where $\varsigma = 0, 1, \cdots, u-1$, and $j \in [2^d]$.

The following theorems ensure the security of the proposed (t, n) VSS scheme.

Theorem 3. Assume that q is a large prime number. The share s_j obtained by the polynomial $f(x)$, has a uniform distribution on \mathbb{Z}_q , $j \in [n]$.

Proof 3. Let A and X be two independent random variables defined on \mathbb{Z}_q . A basic result from the theory of random variables is that if A has a uniform distribution on \mathbb{Z}_q and X has an arbitrary distribution on \mathbb{Z}_q , then $B_1 = A + X \pmod{q}$ and $B_2 = A \cdot X \pmod{q}$ have a uniform distribution on \mathbb{Z}_q , where X is chosen from \mathbb{Z}_q^* in the latter case. If b_1 is chosen uniformly from all possible values of B_1 , the probability of $B_1 = b_1$ is given as:

$$\begin{aligned} Pr[B_1 = b_1] &= Pr[A + X = b_1] \\ &= \sum_{x_j} Pr[A = b_1 - x_j] Pr[X = x_j] \\ &= 1/q \cdot \sum_{x_j} Pr[X = x_j] = 1/q. \end{aligned}$$

Similarly, when b_2 is chosen uniformly from all possible values

of B_2 , we have

$$\begin{aligned} Pr[B_2 = b_2] &= Pr[A \cdot X = b_2] \\ &= \sum_{x_j} Pr[A = b_2 \cdot (x_j)^{-1}] Pr[X = x_j] \\ &= 1/q \cdot \sum_{x_j} Pr[X = x_j] = 1/q. \end{aligned}$$

It can be easily shown that the above argument can be extended to the random polynomial function $f(x)$. Since $a_0, a_1, \cdots, a_{t-1}$ are uniformly distributed on \mathbb{Z}_q and x_j is P_j 's unique nonzero identifier, hence $a_0, a_1 x_j, \cdots, a_{t-1} x_j^{t-1}$ are uniformly distributed on \mathbb{Z}_q . Then, $f(x_j) = a_0 + a_1 x_j + \cdots + a_{t-1} x_j^{t-1}$ is uniformly distributed on \mathbb{Z}_q . Therefore, $s_j = f(x_j)$ is uniformly distributed on \mathbb{Z}_q , that is, s_j has a uniform distribution on \mathbb{Z}_q .

Theorem 4. Under the assumption that H is a secure NAHF, the secret s and some shares s_j cannot be obtained by an attacker from V , $j \in [n]$.

Proof 4. Recall from Definition 3 that an NAHF H is a one-way hash function with quasi-commutativity. Suppose the accumulated item s_j is computed in the j -th iteration of V , thus, $V = H(\cdots H(H(\cdots H(H(k, s), s_1), \cdots, s_j), s_{j+1}), \cdots, s_n)$. Note that $V = H(H(\cdots H(\cdots H(H(k, s), s_1), \cdots, s_{j+1}), \cdots, s_n), s_j) = H(Q, s_j)$, where the first equality holds due to equation (3), and $Q = H(\cdots H(\cdots H(H(k, s), s_1), \cdots, s_{j+1}), \cdots, s_n)$. Furthermore, we have that $H(V, s_j) = V$. We now need to prove that it is hard for the attacker presented with V to find (Q', s_j) such that $V = H(Q', s_j)$. At this point, One-way property of H in Definition 1 ensures that this is computationally infeasible, that is, there is no polynomial P' such that there exists a probabilistic polynomial time algorithm which finds an $s_j \in \mathbb{Z}_q$ such that $V = H(Q', s_j)$ with probability greater than $1/P'(l)$, where Q' is chosen uniformly from $\{0, 1\}^r$. Hence, it is computationally infeasible to find an s_j such that $H(V, s_j) = V$, $j \in [n]$. Similarly, it is computationally infeasible to derive the share s from V .

Theorem 5. In the proposed VSS scheme, any subset of participants of size less than t cannot obtain any information about the secret s .

Proof 5. Here, we consider the worst case, where $t-1$ participants take part in recovering the secret s . Any $t-1$ participants with different identities $x_{j_1}, \cdots, x_{j_{t-1}}$ cannot compute the secret s since they cannot solve the linear system of $(t-1)$ equations and t unknowns: $s_{j_l} = s + a_1 \times x_{j_l} + \cdots + a_{t-1} \times x_{j_l}^{t-1}$, $l \in [t-1]$, which has a degree of freedom, where $a_0 = s$. We can consider the coefficient, a_{t-1} , of the last term in $f(x)$ as a free variable from \mathbb{Z}_q . In this case, the secret s has a unique representation as a linear combination of a_{t-1} and the shares $\{s_{j_1}, \cdots, s_{j_{t-1}}\}$, where a_{t-1} is uniformly distributed over \mathbb{Z}_q . From the proof of Theorem 3, it follows that s has a uniform distribution over \mathbb{Z}_q . Hence, no information about the secret s can be extracted from these $t-1$ shares.

Combining Theorem 3, 4 and 5, we have the following theorem:

Theorem 6. The proposed (t, n) VSS scheme is secure under the assumption that H is a secure NAHF.

TABLE I. THE COMMUNICATION COSTS OF D AND P_j IN THE VSS SCHEMES

	share $ f(x_j) $	verification data	D	P_j
Rajabi-Eslami[3]	$mn_0 p_0 $	$t F(a[i]) =tn_0 p_0 $	$(t+nm)n_0 p_0 $	$(m(t+1)+t)n_0 p_0 $
Feldman [4]	$ q $	$t A_i =t p $	$n q +t p $	$t p +(t+1) q $
proposed scheme	$ q $	$ V =r$	$n q +r$	$(t+1) q +r$

TABLE II. THE COMPUTATION COSTS OF D AND P_j IN THE VSS SCHEMES, WHERE T_o IS THE COMPUTATION TIME FOR THE OPERATION $o \in \{F, H, M(\text{MULTIPLICATION}), e(\text{EXPONENTIATION}), E(\text{EXPONENTIATION ON } R_{p_0}), f(\text{COMPUTING } f(x_j) \text{ ON } R_{p_0})\}$, $pm(\text{POLYNOMIAL MULTIPLICATION ON } R_{p_0})$

	share $ f(x_j) $	verification data	verify $f(x_j)$	get s	D	P_j
Rajabi-Eslami[3]	$T_{f(x_j)}=T_f$	$t T_{F(a[i])}=tT_F$	$(t-1)T_{E+T_F}$	mtT_M	$nT_f + tT_F$	$t(t-1)T_E + mt(T_{pm} + T_M)$
Feldman[4]	$(t-1)T_M$	$t T_{A_i} = tT_e$	tT_e	tT_M	$n(t-1)T_M + tT_e$	$t^2T_e + tT_M$
proposed scheme	$(t-1)T_M$	$T_V = (n+1)T_H$	T_H	tT_M	$n(t-1)T_M + (n+1)T_H$	$(t+1)T_H + tT_M$

V. PERFORMANCE OF PROPOSED VSS SCHEME

In this section we present and discuss the efficiency and scalability for the proposed scheme in Section IV-B. We mainly consider the costs of an extension of the SS scheme to achieve verifiability. By decreasing the number of verification data, we improves on the previous VSS schemes [3, 4]. We estimate the efficiency by counting the number of basic cryptographic operations required in the extension, and also calculate its communication cost. To evaluate the scalability of the proposed scheme, it suffices to show that the costs of each participant remain unchange in the increase in the size of the IOT network (i.e. the number of participants).

Bandwidth is a scarce resource. In a VSS, the communication cost is dominated by the sizes of both verification data and a share. From Table I, we see that in the proposed scheme, the communication costs of the D and P_j are significantly lower than Feldman scheme and Rajabi-Eslami scheme since $|q|$ is much less than $|p|$ and $mn_0|p_0|$ (see Section VI). In the proposed scheme, the verification data V is only a value in \mathbb{Z}_q , so is any share. Specifically, at the share generation phase, the dealer D broadcasts V to participants in P and transmits $s_j=f(x_j)$ to each participant $P_j, j \in [n]$, where $|V| + \sum_{j=1}^n |s_j| = r + n|q|$ bits. Upon receiving V and s_j from D at the share generation phase, each P_j obtains at least $(t - 1)$ different shares s_{j_θ} from the others in P while sending s_j to them at the secret reconstruction phase. Here, $|V| + |s_j| + \sum_{\theta=1}^{t-1} |s_{j_\theta}| + |s_j| = r + (t + 1)|q|$ bits. In the Feldman scheme[4], the verification data included t elements A_0, \dots, A_{t-1} (see Section II) in \mathbb{Z}_p , and the size of each share was the same as the proposed scheme. Therefore, the communication costs at D and P_j were $n|q| + t|p|$ and $t|p| + (t + 1)|q|$ bits, respectively. In the Rajabi-Eslami scheme [3], the verification data was composed of $(n_0 - 1)$ -degree polynomials $F(a[0]), \dots, F(a[t - 1])$ in \mathbb{Z}_{p_0} and each share contained m polynomials in R_{p_0} . Here, the polynomial ring $R_{p_0} = \mathbb{Z}_{p_0}[\alpha]/(\alpha^{n_0} - 1)$, and D_{p_0} was an appropriate subset of “small” elements of $R_{p_0}^{-1}$, where the dimension $m > 1$, the integer module $p_0 \geq 2$ and an error distribution δ . Note that each share $f(x_j)$ and the secret s were respectively

composed of m polynomials in R_{p_0} and D_{p_0} , and $F(a[i])$ was a polynomial in R_{p_0} . Thus, the communication costs at D and P_j were $(t + nm)n_0|p_0|$ and $(m(t + 1) + t)n_0|p_0|$ bits, respectively.

It is generally assumed that in an IoT system, the dealer or server has powerful computing resources and the computing power of IoT devices is limited [22]. Another advantage of proposed scheme is that the computation costs of participants are low since computational requirements are the basic criteria for known IoT devices. For each participant P_j in the proposed scheme, its computation cost is $(t + 1)T_H + tT_M$, where $H(V, s_j), H(V, s_{j_\theta})$ and $H(V, s)$ are respectively computed for verifying $s_j, s_{j_\theta} (\theta \in [t - 1])$ and the recovered s , and t multiplication operation in the Lagrange interpolation formula (1) are performed to recover s . Note that in Rajabi-Eslami scheme [3], $T_f = m(t - 1)T_m$, and $T_F = mT_{pm}$. This was because m polynomials with degree $(t - 1)$ needed to be computed for each $f(x_j)$ in R_{p_0} and $F(X) = \sum_{i=1}^m X_i b_i$. From the experimental results in Section VI, we know that $T_M < T_{pe} < T_H < T_e < T_E$. Table II shows that the computation cost of P_j is the lowest in the proposed scheme. In contrast, the computation cost of D in the proposed scheme, where the time to compute V increases with n , increases due to the use of NAHF H . To compute $s_j = f(x_j)$ for each participant P_j and verification data V, D needs to execute $t - 1$ multiplication operations for $f(x_j)$ and $n + 1$ NAHF operations for $V, j \in [n]$. It means that the computation cost of D is $n(t - 1)T_M + (n + 1)T_H$. Furthermore, the proposed scheme provides the good scalability since the computation and communication costs of P_j remain unchange when the number of participants increases.

VI. SIMULATION EXPERIMENTS

We further evaluate the performance of proposed scheme using simulation experiments. The experiments are conducted on an Intel(R) Core(TM) i7-6700 CPU@3.40 GHz machine with 8.00 GB memory and Windows7 using JDK1.8. We choose to focus on SHA-512 for hashing h in NAHF H with a 128 bit output, where $N = 2^4$ is an upper bound to the number of accumulated items. When $N > 2^4$, we do this by selecting $u = \lceil N/(2^4) \rceil$ different SHA-512 as Remark 1. For Feldman scheme, the parameters p, q were chosen as suggested (see page 21 in [23]), i.e., $|p|=1024$ bits, and $|q|=160$ bits. As for Rajabi-Eslami scheme, according to the LWE parameters for hardware tests (see Table 4 in [24]), the corresponding parameters $(n_0, |p_0|) = (128, 12)$. In addition, let $m = 2$. To give a detailed quantitative analysis, we assume

¹ \mathbb{Z}_{p_0} was the set of integers from 0 to $p_0 - 1$, $\mathbb{Z}_{p_0}[\alpha]$ denoted the set of polynomials with coefficients in \mathbb{Z}_{p_0} . R_{p_0} containd all polynomials of degree less than n_0 with coefficients in \mathbb{Z}_{p_0} , as well as two ring operations, which were polynomial addition and multiplication modulo $\alpha^{n_0} - 1$. Each polynomial in R_{p_0} had n_0 coefficients in \mathbb{Z}_{p_0} , so there was a bijection between R_{p_0} and $\mathbb{Z}_{p_0}^{n_0}$. The compact knapsack problem over R_{p_0} was defined in [21] as follows: given $m = \mathcal{O}(\log_2 n_0)$ elements $b_1, \dots, b_m \in R_{p_0}$ and a target value $c \in R_{p_0}$, found coefficients $X_1, \dots, X_m \in D_{p_0}$ such that $\sum_{i=1}^m X_i b_i = c$.

that participants are MICA2 motes, which work at 8 MHz with a 8-bit processor ATmega128L, and which adopt IEEE 802.15.4 standard. As described in Cao et al. [25], the power level of a MICA2 mote is $U = 3.0$ V, the current draw in active mode is $I = 8.0$ mA, the receiving current draw is $I_r = 10$ mA, the transmitting current draw is $I_t=27$ mA, and the data rate is $r_d = 12.4$ kbps. The cost of receiving (or transmitting) one byte is $E_r = UI_r(8/r_d) = 19.35\mu J$ (or $E_t = UI_t(8/r_d) = 52.26\mu J$). The parameters are fixed in all experiments.

Experiment 1 examines the average time required to run an operation in Table II. With the above parameter settings, we consider the average value of over 160 trials for an operation. The results are as follows: $T_M=0.0022$ milliseconds (ms), $T_H=0.0858$ (ms), $T_e=1.3445$ (ms), $T_{pm}=0.0169$ (ms), $T_E = 1.6071$ (ms). Especially, the average time performing the addition operation is 0.0007ms, which is negligible compared with the others.

Experiment 2 examines the energy consumption of a participant. To compute the electrical energy consumed by a participant during t_p seconds, we apply Joule's law as $E = UI t_p$. From Table II and Table I, we have that $t_p = (t + 1)T_H + tT_M = (t + 1) \times 0.0858 + t \times 0.0022$ (ms), and $(t + 1)|q| + r$ is equal to $4 + 40$ bytes, where transmitting bytes are t and receiving bytes are $t + 20$. For P_j , the energy cost of communication is $20 \times E_t + (t + 20) \times E_r = 19.36t + 1432.2(\mu J)$, and the energy cost of computation is $3 \times 8 \times t_p = 2.112t + 2.0592(\mu J)$. Thus, the energy consumption of P_j is $21.472t + 1434.2592(\mu J) \approx 0.0215t + 1.4343(J)$. We find that the energy cost of computation is cheap compared to data communication. Again, we compare the energy consumption of P_j in the proposed scheme with that of Feldman scheme and Rajabi-Eslami scheme. From Figure 2 (a), it is evident that the energy consumption of P_j increases with the threshold value, but it is relatively stable in the proposed scheme. In particular, the proposed scheme makes P_j have the smallest energy consumption. Note that given a threshold value t , the energy reduction is equal to the difference of the corresponding ordinate values of two schemes in Figure 2(a). For each participant P_j , the results of energy reduction are shown in Figure 2(b). Compared with the Feldman scheme, the energy reduction of the proposed scheme is larger in the Rajabi-Eslami scheme, and the difference increases with t . Furthermore, Figure 2 (c) shows that, compared to the Feldman scheme and Rajabi-Eslami scheme, the energy consumption of P_j in the proposed scheme is respectively reduced by at least 24% and 83% for a secret.

VII. CONCLUSION

In this paper, we give a lightweight (t, n) VSS scheme based on an NAHF [17]. Different from previous VSS schemes, the proposed scheme generates only a NAHF value as the verification data which proves the validity of the shares for all participants. This means that the scheme has less communication cost than previous approaches to achieve the share verification. Another important property is that the computation and communication costs of each participant remain unchanged when the number of its participants increases. It is convenient for building a secure scalable IoT network. At the same time, because the correctness of each share can be efficiently checked, the new participant can verify whether his

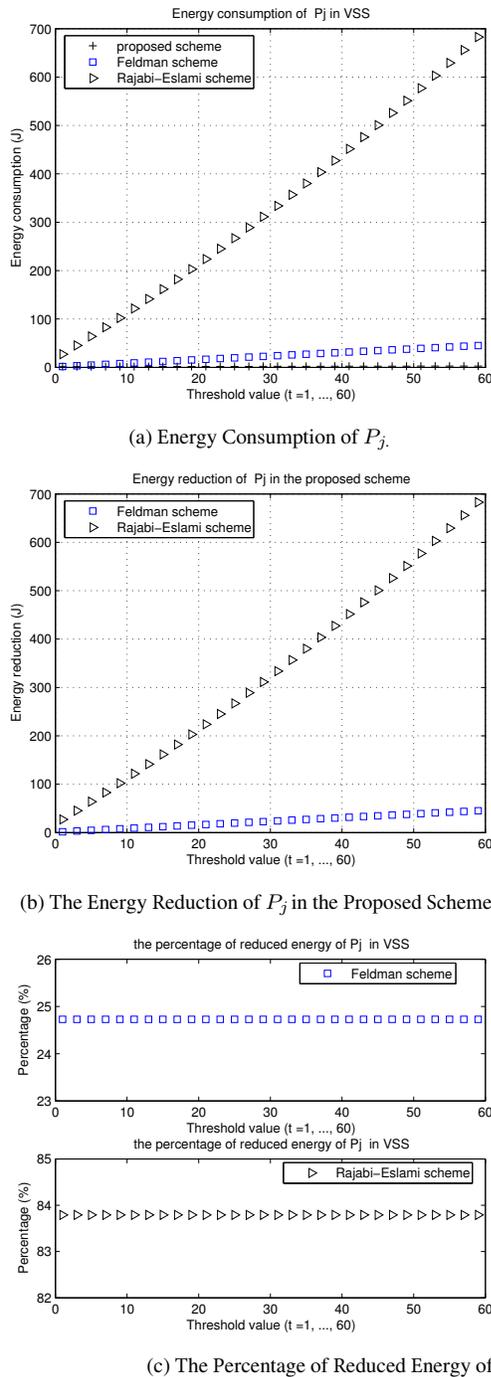


Fig. 2. The Energy Consumption, Energy Reduction and the Percentage of Reduced Energy for P_j in the Proposed Scheme.

share is right or not. Using the proposed scheme, we can make a IoT system more secure and efficient. The presented scheme can be applied to multi-party computation, electronic voting, secure databases and many circumstances.

ACKNOWLEDGMENT

This work is supported in part by National Key Research and Development Plan of China under Grants 2018YFB1003701, in part by the Science and Technology Planning Project of Guangdong under Grants 2021B0101420003, 2020B0909030005, 2020B1212030003, 2020ZDZX3013, and in part by the 22nd batch of Teaching Reform Research Project of Jinan University under Grant 55611518.

REFERENCES

- [1] O. O. Bamasag, K. Youcef-Toumi, Towards continuous authentication in internet of things based on secret sharing scheme, in: Proc. 10st Workshop on Embedded Systems Security, 2015, pp.1-8.
- [2] B. Schoenmakers, A simple publicly verifiable secret sharing scheme and its application to electronic voting, in: Proc. 19th Annu. Int. Cryptol. Conf. Adv. Cryptology, 1999, pp. 148–164.
- [3] B. Rajabi, Z. Eslami, A verifiable threshold secret sharing scheme based on lattices, Information Sciences 501(2019) 655-661.
- [4] P. Feldman, A practical scheme for non-interactive verifiable secret sharing, in: Proc. 28th Annual Symposium on Foundations of Computer Science, 1987, pp. 427–438.
- [5] M. Cafaro, P. Pellè, Space-efficient verifiable secret sharing using polynomial interpolation, IEEE Trans. Cloud Comput. 6(2)(2018) 453-463.
- [6] A. Wander, N. Gura, H. Eberle, V. Gupta, S. C. Shantz, Energy analysis of public-key cryptography for wireless sensor networks, in: PerCom 2005, 2005, pp. 324-328.
- [7] C. Wu, S. Li, Y. Zhang, Key management scheme based on secret sharing for wireless sensor network, in: Proc. 4th Int. Conf. Emerging Intell. Data Web Technol., 2013, pp. 574–578.
- [8] A. Shamir, How to share a secret, Commun. ACM 22(11)(1979) 612–613.
- [9] G. R. Blakley, Safeguarding cryptographic keys. in: Proc. Nat. Comput. Conf., 48, 1979, pp.313-317.
- [10] H. Pirlam, T. Eghlidos: An efficient lattice based multi-stage secret sharing scheme. IEEE Trans. Dependable Secur. Comput. 14(1)(2017) 2-8.
- [11] L. Harn, C. Lin, Strong (n, t, n) verifiable secret sharing scheme, Information Sciences 180(2010) 3059–3064.
- [12] B. Chor, S. Goldwasser, S. Micali, B. Awerbuch, Verifiable secret sharing and achieving simultaneity in the presence of faults (extended abstract), in: Proc. 26th Annual Symposium on Foundations of Computer Science, 1985, pp. 383–395.
- [13] J.C. Benaloh, Secret sharing homomorphisms: keeping shares of a secret secret, in: Proc. on Advances in cryptology–CRYPTO '86, 1987, pp. 251–260 .
- [14] G. Tsaloli, G. Banegas, A. Mitrokotsa, Practical and provably secure distributed aggregation: verifiable additive homomorphic secret sharing. Cryptogr. 4(3): 25 (2020)
- [15] R. Koikara, E.-J. Yoon, A. Paul, Publicly verifiable threshold secret sharing based on three-dimensional-cellular automata. Concurr. Comput. Pract. Exp. 33(22) (2021)
- [16] S. Zhang, Q. Wen, W. Li, H. Zhang, Z.Jin, A multi-user public key encryption with multi-keyword search out of bilinear pairings. Sensors 20(23): 6962 (2020)
- [17] K. Nyberg, Fast accumulated hashing, in: Proc. 3rd Int. Workshop Fast Softw. Encryption, 1996, pp. 83–87.
- [18] J.-J. Huang, W.-S. Juang, C.-I Fan, Y.-F. Tseng, H. Kikuchi, Lightweight authentication scheme with dynamic group members in IoT environments. in: Adjunct Proc. MobiQuitous 2016, 2016, pp. 88-93.
- [19] C.-I Fan, J.-J. Huang, M.-Z. Zhong, R.-H. Hsu, W.-T. Chen, J. Lee, ReHand: Secure region-based fast handover with user anonymity for small cell networks in mobile communications, IEEE Trans. Inf. Forensics Secur. 15(2020) 927-942.
- [20] J. Benaloh, M. de Mare, One-way accumulators: A decentralized alternative to digital signatures, in: Proc. Workshop Theory Appl. Cryptograph. Techn. Adv. Cryptol., 1993, pp. 274–285.
- [21] V. Lyubashevsky, D. Micciancio, Generalized compact knapsacks are collision resistant, in: Proc. 33rd ICALP 2006, 2006, pp. 144-155.
- [22] L. Tawalbeh, F. Muheidat, M. Tawalbeh, M. Quwaidar, IoT Privacy and Security: Challenges and Solutions, Appl. Sci. 10(2020) 4102.
- [23] FIPS PUB 186-2, Digital signature standard (DSS). 2000, January 27. <http://csrc.nist.gov/publications/PubsFIPS.html#fips186-3>.
- [24] N. Göttert, T. Feller, M. Schneider, J. Buchmann, S. A. Huss, On the design of hardware building blocks for modern lattice-based encryption schemes, in: Proc. Int. Workshop CHES, 2012, pp. 512-529.
- [25] X. Cao, W. Kou, L. Dang, B. Zhao, IMBAS: Identity-based multi-user broadcast authentication in wireless sensor networks, Comput. Commun. 31(4)(2008) 659-667.